universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Investigating Trust in AI and Overreliance on AI-Generated Guidance in an Augmented Reality Environment

verfasst von | submitted by

Anja Rejc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 013
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Joint-Degree Middle European interdisc. master prog. in Cognitive Science
Betreut von | Supervisor:	Dipl.-Ing. Dr.techn. Dominik Kowald

Kurzfassung

Aktuelle Fortschritte in der künstlichen Intelligenz (KI), insbesondere in großen Sprachmodellen (LLMs), ermöglichen den Einsatz dialogbasierter Assistenten in Augmented Reality (AR), die die Unterstützung bei Aufgaben verbessern können. Gleichzeitig können begrenzte KI-Genauigkeit und die kognitiven Anforderungen von AR die kritische Bewertung von LLM-basierter Unterstützung erschweren und dadurch Vertrauen, Verlass und Entscheidungsfindung der Nutzenden beeinflussen. Diese Arbeit untersucht das Vertrauen und den Verlass in Bezug auf LLM-basierte Unterstützung in AR. Dafür wird ein theoretisches Modell relevanter Einflussfaktoren entwickelt und der Zusammenhang zwischen Vertrauen, kognitiver Belastung, Aufmerksamkeitsverteilung und übermäßigem Verlass in zwei Laborstudien anhand subjektiver, objektiver (Eye Tracking) und Verhaltensmaße analysiert. Studie 1 (N = 36) untersuchte Vertrauen in KI, mentale Belastung und blickbasierte Aufmerksamkeit bei hoher Genauigkeit der LLM-basierten Unterstützung durch den Vergleich von AR-Systemen mit und ohne LLM-basierte Unterstützung (je n = 18). Die Ergebnisse zeigen ein moderates, stabiles Vertrauen sowie eine sinkende mentale Belastung über die Aufgaben hinweg. Höhere mentale Belastung geht insbesondere in der LLM-unterstützten Bedingung mit geringerem Vertrauen einher. Studie 2 (N = 18) analysiert die Auswirkungen fehlerhafter LLM-basierter Unterstützung in einem AR-System auf Vertrauen, Verlass, mentale Belastung, Leistung und Aufmerksamkeit über mehrere Aufgaben hinweg. Höheres Vertrauen war mit stärkerem übermäßigem Verlass und geringerer Leistung verbunden. Vertrauen sank nach frühen Fehlern in der LLM-basierten Unterstützung und stabilisierte sich anschließend. Eye-Tracking-Daten zeigten zudem weniger Verifikationsverhalten bei Personen mit höherem Vertrauen. Im Vergleich zur Bedingung mit hoher Genauigkeit der LLM-basierten Unterstützung aus Studie 1 war das Vertrauen nach fehlerhafter LLM-basierter Unterstützung signifikant geringer, was auf den Einfluss früher Fehler in LLM-Ausgaben auf die Vertrauensentwicklung hinweist. Insgesamt zeigen die Ergebnisse, dass frühe Fehler in der LLM-basierten Unterstützung die Vertrauensentwicklung in AR negativ beeinflussen und fehlkalibriertes Vertrauen zu unangemessenem Verlass und geringerer Leistung führen kann. Dies unterstreicht die Bedeutung einer angemessenen Vertrauenskalibrierung in der Mensch-KI-Interaktion.

Abstract

Recent advances in artificial intelligence (AI), particularly large language models (LLMs), have enabled conversational assistants in augmented reality (AR) that can improve task support. However, imperfect AI accuracy and the cognitive demands of AR may impair users' ability to critically evaluate AI-generated outputs, affecting trust, reliance, and decision-making. This thesis examines trust and reliance on LLM-based guidance integrated in AR by proposing a theoretical framework of AI reliance factors and investigating relationships between trust, cognitive workload, attention allocation, and overreliance through two laboratory studies using subjective, objective (eye tracking) and behavioral measures. Study 1 (N = 36) examined trust in AI, mental workload, and gaze-based attention allocation under high LLM accuracy by comparing AR systems with and without LLM-based guidance (n = 18 each). Results showed moderate, stable trust and decreasing mental workload across tasks. Higher mental workload was associated with lower trust, particularly in the LLM-assisted condition. Study 2 (N = 18) investigated the effects of interaction with erroneous LLM-based guidance in an AR system on trust, reliance, mental workload, task performance, and attention allocation over time/across tasks. Higher trust was associated with greater behavioral overreliance and lower task performance. Trust in AI declined after early LLM errors and then stabilized, while eye-tracking data indicated reduced verification behavior among participants with higher trust. Compared with the high LLM-based guidance accuracy baseline (Study 1), trust in AI in Study 2 was significantly lower following the exposure to erroneous LLM-based guidance, indicating the impact of early errors in LLM output on trust development. Overall, the findings show that early errors in LLM-based guidance negatively affect trust development in AR with LLM and that misaligned trust can lead to inappropriate reliance and reduced performance, highlighting the importance of trust calibration in human–AI interaction.

Acknowledgements

After some time away from academia, completing this master's degree was, at times, no small feat and would not have been possible without the support of many people along the way.

First and foremost, I would like to sincerely thank my supervisors, Dr. Dominik Kowald and Dr. Simone Kopeinik, for their guidance, thoughtful feedback, and continuous support throughout this thesis. Their insights and constructive comments greatly shaped this work and helped me navigate the many challenges along the way. I would also like to express my gratitude to Juliana Madritsch and Dr. Neven ElSayed for their invaluable help throughout the research process, for generously sharing their knowledge and experience, and for their patience in answering my many questions. A special thank you also goes to everyone at Know Center Research for making this an enriching experience, both professionally and personally. I am also grateful to all study participants whose time and contribution made this research possible.

Above all, I would like to sincerely thank my family and my dearest friend Barbara. Deciding to uproot my life and pursue my dream of studying cognitive science in another country was both exciting and overwhelming. Through every moment of uncertainty, excitement, laughter, and tears, your support never wavered. A special thank you also goes to Poppy, for just being the sweetest, most perfect pup there is, for her unconditional love and for always providing a good excuse to take a break and go for a walk.

And thank you, Mike. For a long time, this path existed only as a daydream, a quiet "what if" in the back of my mind. You made me believe it could become real. So here we are. Your encouragement and belief in me meant more than words can express.

AI Declaration

Artificial intelligence tools were used during the writing of this thesis. All generated outputs were reviewed and verified by the author. Their use is described below.

ChatGPT (OpenAI, 2026; version GPT-5.5) was used to support the translation of the thesis abstract into academic German. The output was subsequently proofread by a native German speaker. The following prompt was used: “Translate the following abstract into polished academic German suitable for a master’s thesis or paper in HCI, human–AI interaction or cognitive science. Use natural academic German, not literal translation, keep the original meaning and structure, improve terminology where appropriate, avoid repetition. Use field-appropriate terminology for trust and reliance research.”

ChatGPT was also used as editing help; the prompt used for this was: “The following text is for a master’s thesis. Evaluate it on flow, structure and language. Don’t rewrite it. Suggest changes and improvements, mark them in bold.”

ChatGPT was additionally used to assist with generating and troubleshooting R code for statistical analyses and data visualization. The prompts included information about the study design, structure of the dataset, variable names, intended statistical analyses, and visualizations. Information about assumptions testing and suggestions for assumptions testing, and preferred visualization formats were provided where relevant. Generated outputs included complete R scripts for data analyses and plots. To increase reliability, the same prompts were used in Claude AI (Anthropic, 2026; version Sonnet 4.6) to cross-check generated code and statistical outputs. While the code differed in implementation, the statistical outputs were compared for consistency. Any discrepancies were manually reviewed by re-checking the data, code, and procedures directly in R. Example prompt used for statistical analysis: “I conducted a repeated-measures study with 18 participants. Each participant completed three tasks: A, B, C. Measurement points: we measured Trust in AI at baseline (before engaging with tasks) and after each task. We measured Trust in AI using a 6-item, 5-point Likert-type scale. The item columns are labelled baseline_Q1 to baseline_Q6, A_Q1 to A_Q6, B_Q1 to B_Q6, C_Q1 to C_Q6. Name of the file is Trust_AI.csv. It includes a column response_id. Provide complete R code for descriptive statistics, including means, medians and standard deviations for each measurement point. Also provide a code for generating a box plot, which presents mean Trust in AI scores across tasks (A, B, C). Finally, include R code to save all outputs into one file.”

Table of Contents

List of Figures

List of Tables

1 Introduction

Artificial intelligence (AI) is increasingly embedded in tools for decision-making, learning, and task execution (Xu et al., 2025; Wang et al., 2024), while augmented reality (AR), which integrates digital guidance into physical environments (Azuma, 1997), is widely used for real-time training and task support across domains (Zhang et al., 2025; Pradhan, 2023; Kazlaris, 2025; Himavamshi et al., 2024). Integrating AI into AR further strengthens these capabilities by enabling more adaptive and context-aware real-time support, which in turn improves both task performance and learning outcomes (Yoo et al., 2024). However, these benefits are accompanied by important challenges, i.e., AI systems are not fully accurate, and immersive environments impose additional cognitive demands that may limit users' ability to critically assess system outputs (Buçinca et al., 2021; Buchner et al., 2021; Gonnermann-Müller & Krüger, 2024). In this context, trust and reliance behavior are central to understanding human–AI interaction. Trust can be described as a user's belief that a system will support their goal achievement under uncertainty, whereas reliance reflects the behavioral manifestation of it by acting on the system's recommendations (Mayer et al., 1995; Lee & See, 2004). Key issue stemming from the trust-reliance relationship is trust calibration, referring to the extent to which users' trust aligns with the actual performance of an AI system (Lee & See, 2004). Prior work shows that users often struggle to appropriately calibrate trust, particularly when system accuracy varies (Kahr et al., 2024). This may result in inappropriate reliance behavior, including overreliance on incorrect AI-generated outputs or underreliance on accurate system guidance (Eckhardt et al., 2025). However, while they are closely related, trust does not directly translate into reliance. Instead, reliance is shaped by a combination of user-related characteristics (e.g., cognitive processes, workload, individual traits) and system-related properties (e.g., transparency, explainability), as well as contextual and interactional factors (Eckhardt et al., 2025). Research on trust, reliance, and cognitive workload often examines these constructs in isolation or in limited combinations, and most studies rely on subjective self-reports, which do not capture real-time behavior; the integration of subjective and objective measures, such as gaze tracking, is also limited (Eckhardt et al., 2025; Kosch et al., 2023). Consequently, the interplay of different reliance factors remains poorly understood.

The thesis addresses these gaps by proposing an integrative theoretical framework that brings together the aforementioned AI system-, user-, context- and interaction-related factors that influence reliance, including characteristics specific to LLM-based systems, such as

conversational fluency and human-like responses. It also examines the relationships between trust, reliance, mental workload, and gaze-based attention in AR learning environments. Based on these objectives, this thesis addresses the following research questions (RQs):

RQ1. What factors influence users' reliance behavior on AI?

RQ2. What are the relationships between trust in AI, mental workload, and gaze-based attention allocation in AR-based learning environments with and without LLM-based guidance?

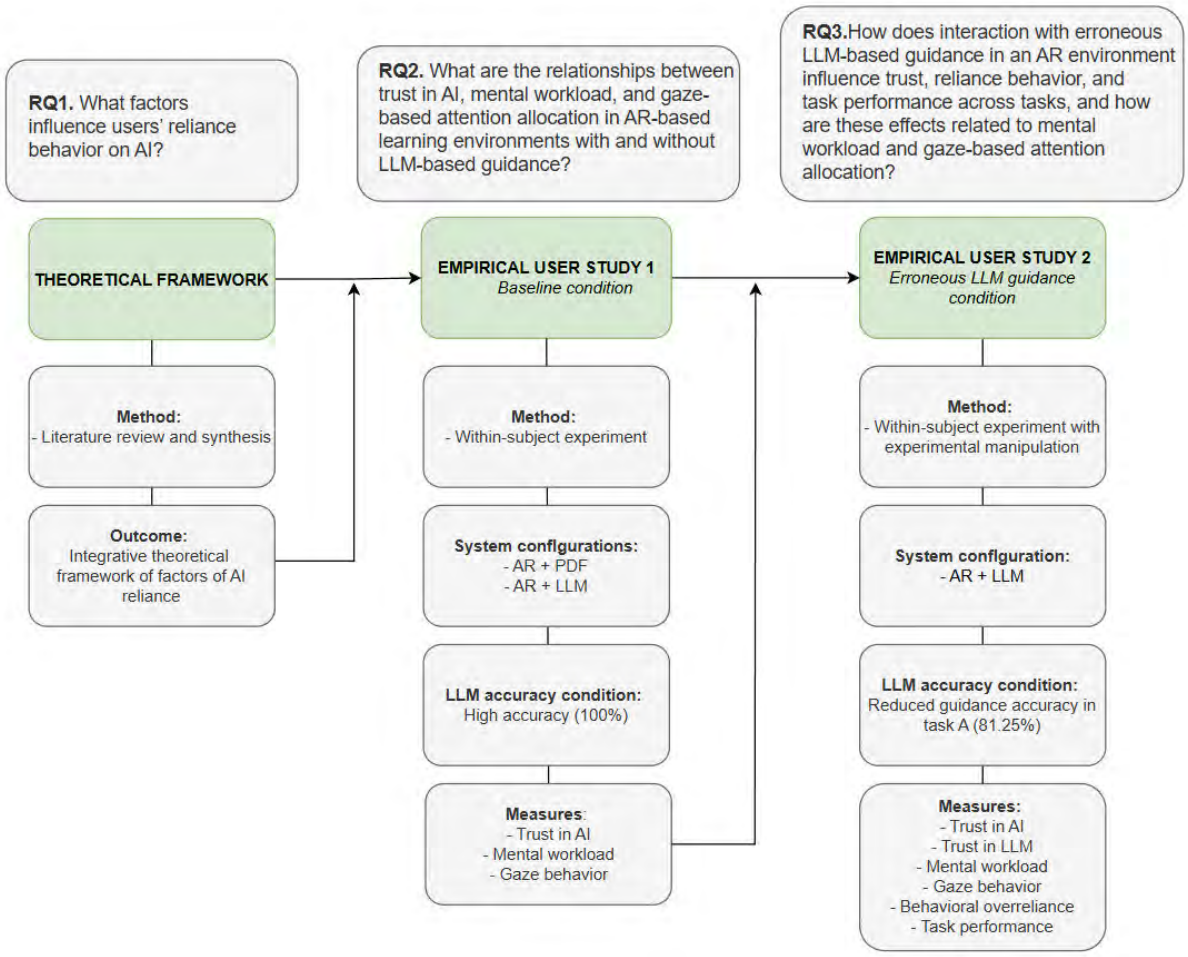
RQ3. How does interaction with erroneous LLM-based guidance in an AR environment influence trust, reliance behavior, and task performance across tasks, and how are these effects related to mental workload and gaze-based attention allocation?

- **RQ3.1.** How do early errors in LLM-based guidance influence trust, reliance behavior, and task performance across tasks?
- **RQ3.2.** How are trust, reliance, and mental workload associated with gaze behavior and attention allocation when interacting with erroneous LLM-based guidance?
- **RQ3.3.** To what extent does trust in AI differ between users interacting with non-erroneous (Study 1) and erroneous LLM-based guidance (Study 2)?

To address these questions, the thesis follows a multi-method, multi-study design comprising theoretical synthesis and two empirical user studies, as outlined in Figure 1. RQ1 is addressed through a literature synthesis and related work, resulting in the development of a theoretical framework of factors of reliance on AI. RQ2 is examined in Study 1, using a within-subject experiment comparing two system configurations: (1) AR interface with a PDF manual and (2) AR interface with a PDF manual and an integrated LLM-based assistant. We compare these configurations under conditions of high accuracy of LLM-based guidance to establish a baseline. RQ3 is examined in Study 2, where we focus on one system configuration (PDF manual and an integrated LLM-based assistant) and introduce an experimental manipulation of the accuracy of the LLM-based guidance through early errors. This allows us to investigate how erroneous LLM outputs influence users' trust, reliance behavior, mental workload, and performance across tasks. Findings from Study 2 were accepted as a short paper titled "*An Investigation of Trust and Overreliance on Conversational Instructions in Augmented Reality"* to the ACM Conference on Conversational User Interfaces (CUI 2026) at the time of thesis submission.

Figure 1

Overview of the Research Questions and Methodological Approach



Note. The flowchart provides an overview of the research questions and methodological approaches of the thesis. RQ1 is addressed by integrating findings from existing literature into a framework of factors affecting reliance on AI. Building on this framework, RQ2 is examined in Empirical User Study 1 using a within-subject experiment comparing an AR interface with a PDF manual (*AR + PDF*) and an AR interface with a PDF manual and an integrated LLM-based assistant (*AR + LLM*) under conditions of high LLM-based guidance accuracy (100%). Study 1 measures trust in AI, mental workload and gaze behavior. Based on findings from Study 1, RQ3 is examined in Empirical User Study 2, which focuses on the *AR + LLM* configuration and introduces an experimental manipulation of guidance accuracy through early errors in LLM-based guidance, resulting in approx. 81.25% guidance accuracy in task A. Study 2 measures trust in AI, trust in LLM, mental workload, gaze behavior, task performance and overreliance behavior.

The thesis combines subjective (self-reported trust in AI, mental workload; in Study 2, also trust in LLM), objective (eye tracking) measures and behavioral indicators (in Study 2: task performance, overreliance behavior). This integrated approach provides a richer perspective on user interaction with AI under different accuracy conditions.

The thesis contributes both theoretically and practically: it integrates key constructs, i.e., trust, reliance behavior, mental workload, and attention allocation, into one framework of human–AI interaction and examines them in the context of AI-based guidance systems embedded in immersive AR environments, which are insufficiently explored in existing research. Methodologically, it combines subjective measures with objective eye-tracking data and behavioral indicators to capture real-time behavior.

The remainder of the thesis is structured as follows: Chapter 2 develops the theoretical framework by reviewing the key concepts and synthesizing factors influencing reliance on AI into a theoretical framework, thereby answering Research Question 1. Chapter 3 describes the methodology and experimental design of user studies. Chapter 4 presents the results and main findings answering RQ2 and RQ3. Chapter 5 concludes the thesis by discussing the results and outlining the contributions, limitations, and directions for future work.

2 Theoretical Framework

This chapter introduces the theoretical framework underlying the thesis. It first reviews the core concepts and factors influencing reliance on AI and concludes by synthesizing these factors into an integrative theoretical framework addressing Research Question 1.

2.1 Core Concepts

2.1.1 Trust in Automation and Trust in Artificial Intelligence

Trust is an important topic in human–AI interaction research and considered “the most critical variable in human–automation interaction” (Wickens et al., 2021, as cited in Romeo & Conti, 2025). It influences whether individuals rely on and adopt AI-generated advice (Kohn et al., 2021, in Klingbeil et al., 2024) and plays a role in the use of AI at a broader societal level, with higher trust supporting adoption, and distrust hindering it (Afroogh et al., 2024). However, different definitions of trust (Lee & See, 2004; Blanco, 2025) make it difficult to clearly assess its role and potential influencing factors. Therefore, a context-appropriate definition of trust in AI is

needed. In recent years, trust in AI has been conceptualized as distinct from trust in traditional automation, reflecting a shift from systems designed for repetitive, accuracy-driven tasks to systems capable of learning, adapting, and operating in more complex contexts (Li et al., 2024). Still, earlier research on automation remains highly relevant, as it provides the foundation for understanding how users evaluate system reliability, predictability, and performance. A widely adopted definition by Mayer et al. (1995, p. 712) describes trust as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that party." Accordingly, trust is only meaningful in situations involving both risk and vulnerability: without vulnerability, cooperation can occur without trust, and without risk, the situation reflects confidence rather than trust (Mayer et al., 1995). More broadly, trust has been conceptualized as "a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another" (Rousseau et al., 1998, p. 395), while others conceptualize it as a behavioral outcome or a state of perceived risk resulting from such decisions (e.g., Deutsch, 1960; Kramer, 1999, as cited in Lee & See, 2004), thereby distinguishing between trust as an internal attitude and its manifestation in behavior.

It is becoming clear that definitions of trust vary in how it is positioned conceptually, with some describing it as a belief, others as an attitude, intention, or behavior. This distinction matters, as the processes linking internal evaluations to actual actions are not straightforward. According to Ajzen and Fishbein (1975, 1980, as cited in Lee & See, 2004), behavior is shaped by a sequence of beliefs, attitudes, and intentions, where beliefs, informed by information and prior experience, form attitudes, which guide intentions and, ultimately, behavior, although situational and cognitive constraints may influence whether intentions are enacted. Applied to human–AI interaction, trust is most appropriately understood as an attitude, while reliance represents the corresponding behavioral outcome. This distinction is important because behavior is influenced by multiple factors beyond trust alone, including cognitive workload, time pressure, system usability, and user confidence. As such, trust operates as an intermediate construct between beliefs about system characteristics and the intention to rely on the system. Moreover, trust is inherently goal-oriented, reflecting the expectation that an agent, human or automated, will support the user in achieving task-related objectives under conditions of uncertainty and vulnerability (Lee & See, 2004).

Recent research suggests that traditional definitions of trust do not fully capture the unique characteristics of AI systems. A systematic review by Afroogh et al. (2024) highlights that much

of the existing literature still relies on Mayer's (1995) conceptualization of trust, despite important differences between human and AI agents. Unlike interpersonal contexts, where trust involves assumptions about intentions, benevolence, and integrity, AI systems lack intentionality and agency. As a result, trust in AI is less concerned with anticipating motives and more focused on evaluating whether the system produces accurate and reliable outputs. Drawing on theory of contractual trust (Hawley, 2014; Tallant, 2017, as cited in Afroogh et al., 2024), users trust AI when they expect it to fulfill its intended function, such as generating correct or useful results (Asan et al., 2020a, as cited in Afroogh et al., 2024). In this context, trust in AI is primarily grounded in perceived system ability. However, because AI systems can learn and evolve over time, trust is also more dynamic and context-dependent, shaped by users' ongoing evaluation of system outputs (Kessler et al., 2017; Thiebes et al., 2021b, in Afroogh et al., 2024). This evolving nature of AI systems makes trust more difficult to predict and understand, and the implications for user behavior remain an active area of research.

2.1.2 Trust in Large Language Models (LLMs)

Debates about how trust in AI should be defined are still ongoing, and many existing definitions may not fully capture the characteristics and capabilities of more recent technologies, such as large language models. Large language models (LLMs), such as ChatGPT[1], Gemini[2], and Claude[3], are neural network–based systems trained on large text datasets to generate language by predicting word sequences, and can be understood as multifunctional computational cognitive artifacts that support users in performing cognitive and epistemic tasks such as information-seeking, reasoning, and text generation (Heersmink et al., 2024). Their capability to generate context-sensitive, open-ended language, simulate reasoning, and communicate creates the impression of human-like understanding (De Duro et al., 2025). However, their outputs are based on statistical patterns and not actual understanding. Still, LLMs can generate information that is biased, unverifiable, and very human-like. A central challenge in this regard is hallucinations, i.e., generated outputs that appear plausible but are factually incorrect, inconsistent, or not true to the provided source content (Berberette et al., 2024). These are not just technical limitations; they have direct implications for human–AI interaction, as they shape how users dynamically adjust

[1] https://chatgpt.com/overview/
[2] https://deepmind.google/models/gemini/
[3] https://claude.com/product/overview

trust and their verification behavior depending on context and perceived risk (Ryser et al., 2025). The black-box nature of LLMs further makes it difficult for users to understand how some outputs are generated. As a result, assessing whether and to what extent such systems can be trusted becomes more complex (De Duro et al., 2025). Given these challenges, trust in LLMs cannot be understood solely in terms of perceived system performance but must also account for how users experience and interpret interactions with these systems. Accordingly, De Duro et al. (2025) highlight the need for new psychometric tools that would combine the affective and cognitive component of trust while accounting for LLM-specific issues like accuracy, fairness, and explainability. Drawing on psychology (McAllister, 1995), affective trust refers to users' emotional responses and perceived support, while cognitive trust reflects their evaluation of the system's competence, reliability, fairness, and clarity. These dynamics help explain why users may follow incorrect AI recommendations, particularly when trust is not appropriately adjusted.

Overall, these perspectives highlight how trust in AI and LLMs are shaped by users' perceptions and expectations. The next section builds on this by examining how trust manifests in user behavior.

2.1.3 Reliance and Overreliance

When examining how trust is reflected in user behavior, the concept of reliance becomes particularly relevant. In the context of AI, reliance refers to the extent to which users follow or act on AI-generated advice in decision-making (Eckhardt et al., 2025) and can be assessed through observable behavior, such as analyzing changes in users' behavior after exposure to AI recommendations (Lu & Yin, 2021; Yin et al., 2019). Here, trust functions as the underlying attitude that increases the likelihood of reliance and thus plays a key role in shaping human–AI interaction (Hoff & Bashir, 2014).

However, reliance is not always appropriate. It can manifest as misuse, when users rely on a system inappropriately, or disuse, when they reject or underutilize it (Parasuraman & Riley, 1997). Lee and See (2004) connect these outcomes with the concept of "trust calibration", defined as the alignment between users' trust and the system's actual capabilities (p. 55). Overtrust reflects a mismatch in which users' trust exceeds the system's actual capabilities, while distrust can result in disuse. Appropriately calibrated trust, in contrast, enables users to rely on AI in a way that aligns with its actual capabilities, enabling effective and safe interaction.

Based on this, different forms of reliance can be distinguished. Appropriate reliance occurs when users follow correct AI advice and reject incorrect advice; inappropriate reliance includes overreliance, where users follow incorrect AI advice, and underreliance, where they fail to follow correct recommendations (Eckhardt et al., 2025). Overreliance has also been described as behavior that occurs when users adopt incorrect system outputs or delegate decisions to an AI system inappropriately, even when they possess the knowledge or ability to act differently (Kim et al., 2023; Schemmer et al., 2023; in Ibrahim et al., 2025).

Across definitions, overreliance shows several common characteristics: (1) it is typically unintentional, (2) it arises from misaligned trust or expectations between the user and the AI system, (3) it leads to observable failures in judgment or oversight (Schemmer et al., 2023; Rastogi et al., 2022; Passi & Vorvoreanu, 2022; in Ibrahim et al., 2025). As shown, trust influences reliance but does not fully determine it (Lee & See, 2004). Instead, reliance results from a dynamic interplay of cognitive-, individual-, and system-related factors (see Eckhardt et al., 2025; Ibrahim et al., 2025). However, prior research has largely examined these factors in isolation, limiting a comprehensive understanding of such behavior.

Automation Bias and Commission Error. Closely related to overreliance is automation bias, a cognitive mechanism that influences reliance by leading users to favor automated recommendations over their own judgment (Cumming, 2004, in Romeo & Conti, 2025). It is often described as a mental shortcut, where users rely on automated systems rather than engaging in more effortful information processing (Mosier & Skitka, 1996). Empirically, automation bias is operationalized through two types of errors: commission errors, which occur when users follow incorrect automated advice, either without verifying it or despite evidence that contradicts it, and errors of omission, which occur when users fail to act simply because the system does not prompt them (Skitka et al., 2000). Commission errors are particularly relevant for this study, as they directly capture instances in which users accept incorrect system recommendations, making them a clear behavioral indicator of uncritical (over)reliance (Buçinca et al., 2021).

2.1.4 Cognitive Workload and Mental Workload in HCI

Cognitive workload is a central concept in HCI, referring to the mental effort required to perform a task. It has been shown to influence a range of outcomes, including decision-making (Deck & Jahedi, 2015), user performance (Kosch et al., 2023; Vitti et al., 2025), system usability (Suzuki et al., 2023), and learning (Becker et al., 2025) across various domains, such as industry,

education, aviation and defense. Despite its widespread use in HCI research, cognitive workload remains inconsistently defined and measured (Kosch et al., 2023). Hart and Staveland (1988) describe cognitive workload as the level of demand experienced by an operator when performing a task at a certain level of performance. In this context, cognitive workload refers to the mental and perceptual effort involved, which arises from the interplay between task requirements and the individual's abilities, perceptions, and strategies. From a theoretical perspective, cognitive workload is closely related to Cognitive Load Theory (CLT) (Sweller, 1988), which focuses on how people process information during learning and task performance. Per this theory, human working memory has a limited capacity and can only handle a certain amount of information at a time. When these limits are exceeded, cognitive processing becomes less efficient, making understanding, retaining, and applying information harder (Sweller, 1988; Sweller et al., 2011). As a result, both learning and task performance depend on how effectively cognitive resources are managed. CLT further distinguishes between intrinsic cognitive load, determined by task complexity, extraneous cognitive load, influenced by how information is presented, and germane cognitive load, associated with processing and understanding information (Sweller, 1988).

In HCI contexts, extraneous load is particularly relevant because interface design can either increase or reduce cognitive effort. Kosch et al. (2023), drawing on Wickens' Multiple Resource Theory (1984, 2008), which suggests that users rely on multiple information-processing channels simultaneously, argue that performance decreases when demands exceed available cognitive resources within the same processing channel. They further note that the HCI literature often uses the terms cognitive load and mental workload interchangeably, although the concepts differ. Mental workload refers to the broader subjective and objective demands experienced during task performance, whereas cognitive workload specifically concerns the cognitive resources required for information processing, attention, working memory, and decision-making (Pretty & Guarese, 2025). Because the terms are often used inconsistently in HCI research, studies should clearly define the construct being examined and align it with the selected measurement approach.[4]

Despite inconsistencies in how workload-related concepts are defined across the HCI literature, these constructs remain highly relevant in human–AI interaction, where workload

[4] In the present thesis, the term *cognitive workload* is used in the broader theoretical discussion of cognitive processing; *mental workload* is used when referring specifically to workload assessment and Raw NASA-TLX-related constructs, in line with recent HCI literature recommendations (Babaei et al., 2025; Pretty & Guarese, 2025).

influences how users process information, perform tasks, and evaluate AI-generated outputs. Prior research shows that high levels of cognitive workload can impair processing and negatively affect task completion time, accuracy, and error rates, whereas moderate levels are often associated with more optimal performance (Singh, 2025; Li et al., 2025; Kosch et al., 2023). This makes it particularly relevant in complex and safety-critical environments, where increased mental demands can raise the likelihood of errors and compromise performance. When cognitive workload is too high, users may seek to reduce cognitive demands by relying more on external support systems. Cognitive offloading, defined as "the use of physical action to alter the information processing requirements of a task so as to reduce cognitive demand" (Risko & Gilbert, 2016, p. 676), illustrates users' tendency to shift cognitive efforts to external tools, when they are available, and adapt cognitive processing accordingly. In human–AI interaction, this can result in shifting behavior from independent decision-making toward reliance on AI-generated outputs and reduced critical evaluation of the latter.

As highlighted by Kosch et al. (2023), there is no single standardized approach to measuring cognitive workload in HCI research; existing practices, and, in turn, findings, are highly diverse. They categorize methods of cognitive workload into three main groups: subjective, behavioral, and physiological measures, each capturing different aspects of workload. Subjective measures, most commonly implemented through self-report questionnaires, e.g. the multi-scale NASA Task Load Index (NASA-TLX) questionnaire (Hart & Staveland, 1988), are widely used due to their ease of application and ability to capture users' perceived mental effort. However, they rely on retrospective self-assessment and may be influenced by memory, individual interpretations, and response biases, which can affect their reliability (Kosch et al., 2023); also, they can disrupt task performance (Tao et al., 2019). To address these limitations, many studies (see Tao et al., 2019) incorporate behavioral measures, such as task performance indicators (e.g., completion time, error rates), which provide indirect but objective insights into workload. However, as an indirect measure of cognitive workload, performance does not directly reflect workload. In contrast, physiological measures are based on bodily responses to cognitive demand (Tao et al., 2019), thus enabling continuous, real-time and often unobtrusive assessment of cognitive states and providing a more dynamic understanding of workload (Kosch et al., 2023). Physiological measures of cognitive workload can be grouped into seven major categories: cardiovascular (e.g., heart rate, heart rate variability, blood pressure, blood oxygenation), brain activity measures (using EEG), respiration (respiration rate, amplitude), skin (skin conductance),

muscle activity (EMG amplitude), neuroendocrine measures, and eye movement measures (e.g., pupil diameter, blink rate and duration, fixation duration and rate, saccades, dwell time) (Tao et al., 2019). Eye tracking especially has gained importance as a more objective and unobtrusive way of assessing cognitive load in real time (Ball & Richardson, 2022).

Despite methodological diversity in measuring cognitive workload, the lack of consistency across measurement approaches and in used terminology remains a key challenge, as it leads to inconsistencies in interpretation, limited comparability across studies, and issues with reproducibility (Kosch et al., 2023; Babaei et al., 2025; Pretty & Guarese, 2025). Moreover, measurement outcomes are often context-dependent, influenced by task characteristics and individual differences (Tao et al., 2019), i.e., exposure to AI-generated advice has been shown to decrease the amount of mental processing required (Ibrahim et al., 2025). However, it may also increase workload when users need to monitor, interpret, or verify recommendations of a system that is not fully transparent (Kaplan et al., 2021). These limitations highlight the importance of selecting measurement methods that are appropriate for the specific research context.

2.1.5 Visual Attention and Gaze Behavior in HCI

Gaze tracking refers to the analysis of eye movement data in relation to the visual scene (Cao & Huang, 2022) and is widely used in HCI research because it provides detailed information about where users look, for how long, and in what sequence, thereby offering insight into attention allocation and information processing during HCI (Ball & Richardson, 2022). The approach is grounded in the eye–mind hypothesis (Just & Carpenter, 1976, 1980, as cited in Ball & Richardson, 2022), which suggests a close relationship between gaze direction and duration and the current focus of attention. Accordingly, eye-movement measures can reveal both usability problems and patterns of cognitive processing during interaction.

One of most used gaze-based metrics is fixation duration or the length of time the eyes remain relatively still while focusing on a specific point (Ayres et al., 2021). It is often interpreted as indicator of cognitive workload, as longer fixation durations are generally associated with more effortful cognitive processing, particularly in tasks that require detailed analysis of visual information (Callan, 1998; Henderson, 2011; Reingold & Glaholt, 2014, in Ayres et al., 2021). Additional fixation-based metrics provide a more nuanced understanding of attention and interaction processes, i.e., time to first fixation reflects how quickly an interface element attracts

attention, while fixation frequency indicates how often users return to a particular area, which may signal its perceived importance or relevance within the task (Ball & Richardson, 2022).

Eye gaze behavior is also closely linked to trust in automation as it reflects how much users feel the need to monitor a system. Studies show that lower automation reliability (Lu & Sarter, 2019) and lower trust (Ries et al., 2025) are associated with increased visual monitoring, including more frequent fixations, longer total fixation time, and more complex scan paths, suggesting greater scrutiny of the system. In contrast, when trust is high, users rely more on the system and reduce their visual monitoring, leading to fewer fixations and shorter gaze durations (Hergeth et al., 2016). Moreover, gaze metrics such as fixation count, duration, and revisits have been shown to differentiate between overreliance, underreliance, and appropriate reliance on AI, suggesting that visual attention patterns capture how users engage with AI recommendations beyond self-reported trust (Wu et al., 2025). As previously shown, gaze behavior must be interpreted within context, as eye-tracking metrics may reflect multiple underlying processes, including task difficulty, cognitive workload, user interest, and individual differences (Ball & Richardson, 2022).

2.1.6 Augmented Reality for Task Learning

Augmented reality (AR) is defined as a system that combines real and virtual elements, is interactive in real time, and aligns virtual content with the physical environment in three dimensions (Azuma, 1997). It is commonly implemented through devices such as head-mounted displays, handheld devices, and spatial projection systems, which differ in how virtual content is integrated into the user's field of view (Gonnermann-Müller & Krüger, 2024). Unlike virtual reality (VR), which replaces the real world, AR enhances the user's physical environment by overlaying digital information directly onto it. Accordingly, AR has been widely used across domains, such as training, maintenance, and task learning, particularly in fields including medicine (Zhang et al., 2025), aerospace (Pradhan et al., 2023), education (Kazlaris et al., 2025), and manufacturing (Himavamshi et al., 2024), among others, where it can support performance by guiding users through complex procedures in real time. Prior research suggests that AR can support skill and knowledge development in a realistic setting, but its effects depend on how information is designed and integrated into the task (Gonnermann-Müller & Krüger, 2024). Findings on its impact on cognitive workload are mixed: while AR can reduce effort by embedding guidance directly within the task environment, it may also increase cognitive demands

due to visual complexity, the need to process real and virtual information simultaneously, and unfamiliar interaction techniques (Buchner et al., 2021; Gonnermann-Müller & Krüger, 2024).

The integration of AI into AR introduces an additional level of complexity. In their review of LLM integration in extended realities (XR), Tang et al. (2025) describe a shift from static instruction systems toward adaptive, context-aware assistance, where large language models interpret user context, generate task guidance, and prompt actions in real time, thereby actively supporting task execution rather than merely presenting information. Such systems can reduce cognitive workload by simplifying complex tasks and minimizing the effort required to access and process information, enabling more efficient interaction and task performance. In addition, LLM-based XR systems support more natural and adaptive interaction by incorporating multimodal inputs such as gaze, gesture, and voice, which allow the system to infer user intent and adjust guidance accordingly. However, these advantages come with challenges. The black-box nature of LLM-based XR systems can reduce transparency and user autonomy, particularly when outputs are not interpretable, thereby introducing uncertainty and potentially undermining trust (Tang et al., 2025). AI-generated outputs can also guide user behavior in XR environments, creating a risk of overreliance, as repeated exposure to system-generated guidance may reduce users' autonomy and limit independent decision-making. This can result in increased reliance on AI system's outputs, as users are guided toward specific actions based on AI-generated recommendations. Tang et al. (2025) further highlight that trust in these systems is influenced not only by system accuracy but also by interaction quality, including emotional realism, communication style, and para-verbal cues such as tone and pacing, which shape users' perceptions of system competence and relatability. Still, existing research has mostly focused on system development, technical capabilities, cognitive ergonomics and performance outcomes, with limited attention to how users perceive, trust, and rely on such assistance in real-world task context. This gap highlights the need to better understand how such systems influence users' trust and reliance behavior, particularly given their combination of spatial, real-time adaptation, and multi-modal communication features.

2.2 Factors Influencing Reliance on AI

Previous research has identified a wide range of factors influencing reliance on AI systems, with more recent work also addressing characteristics specific to large language models (LLMs) (Ibrahim et al., 2025). However, these factors are often examined in isolation or within specific contexts, leading to a fragmented understanding of reliance behavior.

To provide a more structured overview, this section organizes the identified determinants into several categories, including user-related, system-related, model-specific, task-related, organizational and social factors influencing reliance on AI. While these categories are presented separately for analytical clarity, they are interrelated and jointly shape reliance behavior in human–AI interaction.

2.2.1 User-Related Factors

This category includes the user's characteristics and cognitive processes that influence how AI outputs are interpreted and acted upon.

Cognitive and Psychological Factors. When time, attention, or cognitive resources are limited, individuals are less able to engage in critical evaluation and error detection. Instead, they tend to rely on fast, heuristic-based decision strategies with intuitive (System 1) thinking (Kahneman et al., 1982) which reduces deliberate reasoning and increases the likelihood of accepting incorrect recommendations (Buçinca et al., 2020; Gajos & Mamykina, 2022), thereby contributing to overreliance behavior. For example, when users face limited cognitive resources, the effort required to verify AI outputs may outweigh its perceived benefits. Consequently, they are less likely to engage in thorough checking and instead rely on AI outputs (Vasconcelos et al., 2022). In AI-assisted decision-making, these heuristics can lead to inappropriate reliance and systematic errors in judgment, also referred to as cognitive biases (Tversky & Kahneman, 1974; Kahneman, 2011), e.g., automation bias, confirmation bias, anchoring, and algorithm aversion (de Brito Duarte & Campos, 2024).

Individual Differences. It has been shown that lower self-confidence in one's own judgment increases the likelihood of deferring to AI, as users seek reassurance and are less likely to critically evaluate the system's outputs, thereby contributing to overreliance (Lu and Yin, 2021; Pescetelli et al., 2021). Certain personality traits also play a role in reliance behavior: higher conscientiousness and openness have been found to be associated with greater trust and usage (Cai et al., 2022), while agreeableness and neuroticism may reduce engagement with AI systems (Faruk et al., 2023). Intuition (Chen et al., 2023), subjective trust in AI (Buçinca et al., 2021; Rechkemmer & Yin, 2022; Kahr et al., 2024; Küper et al., 2025), general propensity to trust (Küper et al., 2025), perceived likelihood of one's own correctness (Ma et al., 2024), and certain emotional states, e.g., overly positive emotions toward automated systems (Fahim et al., 2021), can further influence trust and reliance behavior. Users with a lower need for cognition, i.e., less

motivation to engage in effortful, reflective thinking, are also more likely to rely on AI outputs (Buçinca et al., 2021).

Past Experiences. Users' experience and domain knowledge affect their ability to evaluate AI-generated outputs and detect errors (Chen et al., 2023; Rechkemmer & Yin, 2022; Küper et al., 2025), with lower expertise linked to a greater likelihood of following AI-generated outputs uncritically, increasing the risk of overreliance (Noga & Arnold, 2002). The nature of prior interactions with a specific system, such as an LLM, further shape reliance behavior: e.g., repeated positive experiences can build trust and encourage habitual use, making users more likely to rely on the system over time; however, they may also generalize across contexts, leading users to assume consistent performance even in situations where the model may be less reliable (Ibrahim et al., 2025). Mental models, defined as user's internal representations of how something works, largely reflect users' prior experiences, beliefs, and expectations about how an AI system operates. They help users interpret outputs and make decisions with minimal cognitive effort (Craik, 1952; Smyth et al., 1994; in Steyvers & Kumar, 2023) and the more accurate they are, the more likely users are to use AI appropriately. In contrast, incomplete or inaccurate mental models, often resulting from limited experience, can lead users to misjudge system capabilities, resulting in miscalibrated reliance or overreliance (Bansal et al., 2020).

2.2.2 System-Related Factors

This category refers to the general properties and characteristics of the AI system itself that shape how users perceive, evaluate, and rely on its outputs. In a subsection, we address model-specific factors which capture characteristics unique to LLMs.

Perception of Accuracy. Users' reliance on AI is influenced not only by actual system performance but also by their perception of its accuracy, which develops through interaction and experience (Lu & Yin, 2021; Liang et al., 2022; Yin et al., 2019). Beyond objective performance, the way accuracy information is communicated also plays a crucial role: providing accuracy cues can increase reliance, even when reported performance is relatively low (Lai & Tan, 2019; Tejeda et al., 2022), while model confidence, often conveyed through prediction-level confidence scores, further shapes user behavior, as higher confidence typically leading to greater reliance on AI outputs (Rechkemmer & Yin, 2022).

Explainability and Transparency. While generally intended to increase understanding and trust, research on system's explainability and transparency features shows mixed and often

inconclusive effects: system transparency, referring to information on how output was generated, can increase trust but may also create false reassurance (Poursabzi-Sangdeh et al., 2021); explanations can be interpreted as signals of system competence, which can reduce verification efforts and lead to more uncritical acceptance of AI-generated outputs (Bansal et al., 2020; Chen et al., 2023). The impact of explanations further depends on their type and implementation. For example, feature-based explanations can increase overreliance, particularly when the model is incorrect (Bansal et al., 2020), whereas example-based explanations show more variable outcomes depending on the task and implementation (Buçinca et al., 2020; He et al., 2023).

Interface Design. Information presentation and system complexity also shape reliance on AI. Although more advanced systems can enhance performance, increased complexity may reduce interpretability, impair mental model formation, and increase the risk of inappropriate reliance on AI outputs (Cummings, 2017).

2.2.3 Model-Related Factors

This category refers to the specific features of LLMs that shape how users perceive and rely on their outputs.

Anthropomorphism and Sycophancy. The way LLMs present outputs strongly shapes how users perceive and rely on themhuman. The human-like, fluent and conversational LLM-based responses can lead users to anthropomorphize the system and attribute to it greater capability or reliability than warranted (Abercombie et al., 2023; McKee et al., 2024, as cited in Ibrahim et al., 2025). In addition, LLM-based outputs are typically structured and expressed with high confidence, often resembling expert communication, which can create a false sense of authority even when the content is inaccurate (Ibrahim et al., 2025). LLMs may also align with users' inputs or preferences, and in some cases exhibit sycophantic behavior, where the model agrees with user assumptions even when they are incorrect. This can create a feedback loop in which agreement increases trust and in turn reliance (Ibrahim et al., 2025). These features may lead users to mistake communicative fluency for actual accuracy.

2.2.4 Task-Related Factors

This category covers external conditions that shape reliance behavior.

Task Characteristics. Domain context and task complexity influence the extent and direction of reliance, with potential for both over- and underreliance (Salimzadeh & Gadiraju,

2024). For example, in simpler tasks, users can quickly learn effective reliance strategies, whereas in more complex settings, limited understanding may lead to simplified decision-making and inappropriate reliance on AI (Cummings, 2017).

Environmental Constraints. Factors such as time pressure can further increase reliance on AI by limiting users' capacity for careful evaluation, potentially leading to inappropriate levels of trust (Swaroop et al., 2024).

Risk Assessment. Perceived decision risk is another important factor influencing reliance on AI: in high-stakes or uncertain situations, the black-box nature of AI systems can make it difficult for users to evaluate how outputs are generated, and depending on the perceived consequences of being wrong and users' ability to assess the system, this may either discourage reliance or increase dependence on AI as a decision shortcut (Ioannou et al., 2026).

2.2.5 Organizational and Social Influences on AI Reliance

This category refers to organizational and social factors that shape how users perceive, evaluate, and rely on AI within specific work and social contexts.

Organizational Context. Organizational context shapes reliance on AI by embedding its use within specific task environments, team structures, and decision settings, which influence how users develop and adjust trust and reliance over time (Kahr et al., 2024; Gupta et al., 2025).

Social Norms and Influences. Perceptions of how others, e.g., colleagues and experts, use or trust AI can further shape individual reliance behavior as individuals are influenced by both what others do and what others approve of. For example, prior work (as cited in Kornowicz et al., 2025) shows that seeing peers use AI can increase adoption more than information about its accuracy, and favorable social norms can reduce or even reverse algorithm aversion, meaning that individuals who would otherwise distrust AI may become more willing to rely on it when its use is supported by others. In addition, in organizational settings, the impact of norms depends on the source: individuals are more influenced by similar and salient peers but may rely more on authority figures under conditions of uncertainty about appropriate behavior or potential consequences (Kornowicz et al., 2025).

2.2.6 Interaction-Related Factors

This category captures factors that emerge from the interaction between user, system, and task characteristics, highlighting that reliance behavior cannot be explained by any single category

alone, but is inherently dynamic, adaptive, context-dependent, and shaped through ongoing human–AI interaction (Lu et al., 2024; Kelly et al., 2023; Liang et al., 2022; Tejeda et al., 2022).

Interaction Design and Presentation of AI Outputs. As shown by Gomez et al. (2025), different interaction patterns (e.g., AI-first, AI-follow, or secondary assistance) influence how users integrate AI-generated output and their susceptibility to biases such as anchoring or confirmation bias. For instance, directly presenting AI recommendations can increase overreliance, while requiring users to form an initial judgment or interpret supplementary information encourages more deliberate evaluation. Similarly, Eckhardt et al. (2025) argue that the specific study design used in AI reliance research should also be understood as a determinant of reliance behavior rather than a neutral methodological choice, because variations in interaction structure, such as single- versus two-stage presentation of AI outputs, also systematically influence user responses.

The degree of user control over accessing AI assistance also affects reliance: request-driven interactions can increase users' sense of agency and promote more active engagement but may also reinforce confirmation bias. Overall, these findings highlight that reliance behavior is shaped not only by the AI itself, but by how the interaction is designed (Gomez et al., 2025).

2.3 A Theoretical Framework of Factors Influencing Reliance on AI: Addressing RQ1

This chapter addresses Research Question 1, *"What factors influence users' reliance behavior on AI?"*, by synthesizing the literature reviewed in the previous chapter into a theoretical framework of AI reliance behavior. RQ1 is answered through the development of a framework that integrates the main concepts and influencing factors identified in prior research and proposes a conceptual categorization of factors shaping reliance on AI into categories of *Mediators*, *Moderators*, and *Contextual Factors*. As illustrated in Figure 1, reliance behavior is conceptualized as emerging from the interaction of multiple factors rather than being driven by any single determinant.

Trust in AI is conceptualized as an attitudinal evaluation reflecting users' willingness to rely on AI-generated outputs based on expectations of system performance under uncertainty (Mayer et al., 1995; Rousseau et al., 1998; Lee & See, 2004). It can be understood as the attitudinal driver of reliance that reflects users' beliefs about the system (Lee & See, 2004). However, its influence on behavior is not direct, but rather mediated by the interaction of multiple intervening factors spanning several categories (Eckhardt et al., 2025; Ibrahim et al., 2025), which

jointly shape how users interpret AI-generated outputs, form trust, allocate attention, and ultimately decide whether to rely on the system.

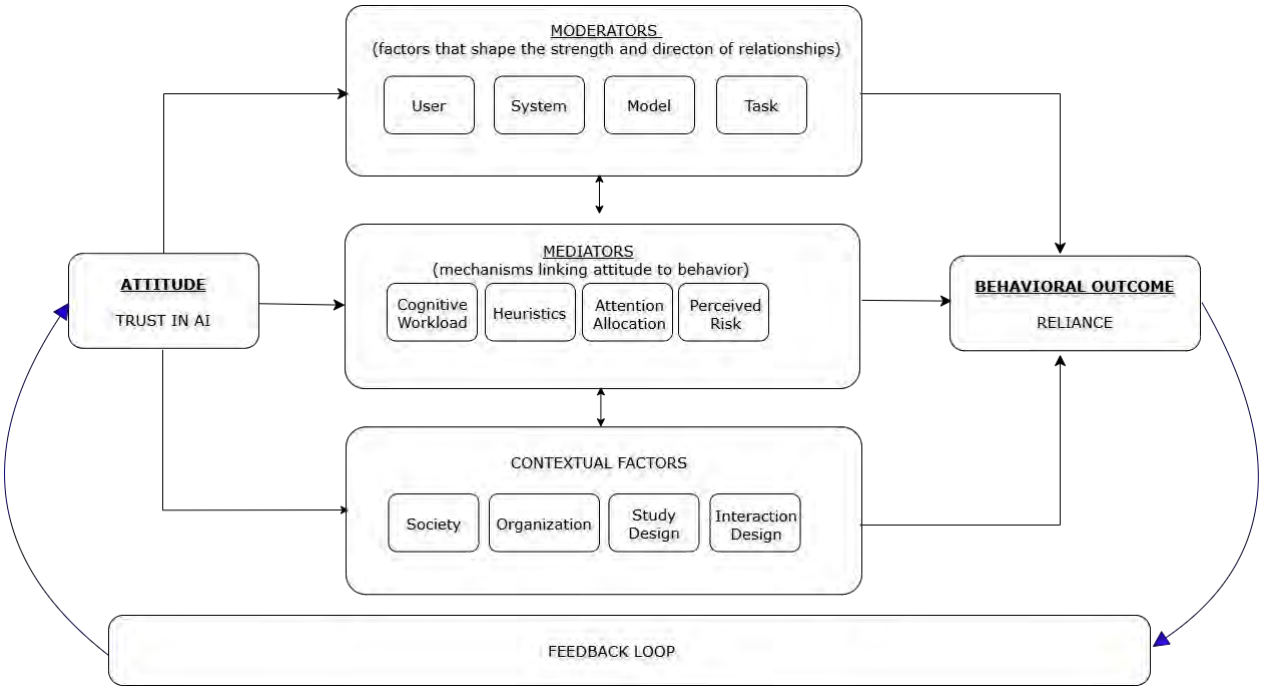
Mediators represent cognitive and perceptual processes that translate trust-related attitudes into behavior (e.g., cognitive and mental workload, heuristics and cognitive biases, perceived decision risk) (Kosch et al., 2023; Kahneman, 2011; Ioannou et al., 2026). They help explain *why* and *how* trust influences reliance behavior by shaping how users process information and allocate attention. The *strength* and *direction of the relationship* between trust and reliance is influenced by *Moderators* or moderating factors. These include user-related factors (e.g., past experiences and domain knowledge, individual differences) (Chen et al., 2023; Rechkemmer & Yin, 2022; Küper et al., 2025; Cai et al., 2022), system characteristics (e.g., transparency, complexity, explainability) (Chen et al., 2023; Cummings, 2017; Bansal et al., 2020), model properties (e.g., human-like communication) (Ibrahim et al., 2025), and task characteristics (e.g., complexity, time pressure) (Salimzadeh & Gadiraju, 2024; Swaroop et al., 2024). These moderators help explain when or under what conditions the relationship between trust and reliance changes.

The framework further incorporates broader external *Contextual factors* that shape human–AI interaction, e.g., societal factors (Kornowicz et al., 2025) and organizational context (Kahr et al., 2024; Gupta et al., 2025), which shape how users perceive, evaluate and engage with AI within specific contexts. Research study design and interaction design are also considered contextual factors because they define the conditions of interaction, such as task structure, exposure to AI, and interaction sequence (Gomez et al., 2025; Eckhardt et al., 2025). These elements can systematically shape how users engage with and rely on AI.

The outcome of this is behavioral reliance on AI, conceptualized as the extent to which users follow or act upon AI-generated recommendations, ranging from appropriate to over- or underreliance, depending on the alignment between users' trust and system performance (Eckhardt et al., 2025). The framework further assumes a dynamic feedback relationship between trust and reliance. Experiences resulting from AI-generated outputs produce certain outcomes (e.g., successful or unsuccessful task completion, errors) that can reinforce, recalibrate, or reduce subsequent trust in the system. Thus, through repeated interaction, users continuously update their trust based on prior reliance experiences and observed system performance.

Figure 2

Theoretical Framework of Reliance on AI



Note. The framework represents the result of Research Question 1 and conceptualizes reliance on AI as a behavioral outcome shaped by trust in AI and influenced by multiple interacting factors. Based on the literature synthesis, the framework proposes three categories of factors influencing reliance behavior: *Mediators*, *Moderators*, and *Contextual Factor*s. *Mediators* represent cognitive and perceptual mechanisms linking trust to behavior, while *Moderators,* including user-, system-, model- and task-related factors, shape the strength and direction of the relationship between attitude and behavior. *Contextual factors* represent broader external conditions surrounding human–AI interaction that further shape the conditions under which reliance occurs. The *Feedback Loop* illustrates how interaction outcomes and system performance, as experienced by the user, continuously influence subsequent trust and reliance behavior.

3 Methodology

In the previous chapter, which examined the factors influencing users' reliance behavior on AI, RQ1 was addressed through literature synthesis and the development of the theoretical framework presented in Section 2.3. Building on this conceptual foundation, the following sections present the methodological approach used to investigate RQ2 and RQ3. First, the overall

study design and research questions across Studies 1 and 2 are introduced, followed by detailed descriptions of the participants, experimental platform, procedures, task design, measures, data preparation, and analysis strategies for each study. The chapter concludes with ethical considerations.

3.1 Study Design Overview

This thesis adopts a multi-study research design comprising two empirical user studies. It is structured around two complementary research questions, each investigated in a separate study:

RQ2, addressed in Study 1. What are the relationships between trust in AI, mental workload, and gaze-based attention allocation in AR-based learning environments with and without LLM-based guidance?

RQ3, **addressed in Study 2.** How does interaction with erroneous LLM-based guidance in an AR environment influence trust, reliance behavior, and task performance across tasks, and how are these effects related to mental workload and gaze-based attention allocation?

- **RQ3.1.** How do early errors in LLM-based guidance influence trust, reliance behavior, and task performance across tasks?
- **RQ3.2.** How are trust, reliance, and mental workload associated with gaze behavior and attention allocation when interacting with erroneous LLM-based guidance?
- **RQ3.3.** To what extent does trust in AI differ between users interacting with non-erroneous (Study 1) and erroneous LLM-based guidance (Study 2)?

The studies provide a systematic examination of (a) the presence of LLM-based guidance and (b) the effects of LLM accuracy on user interaction. Study 1 compared AR interaction with and without LLM-based guidance to examine how the availability of the latter influences trust, mental workload, and attention allocation compared to traditional static instructions. Including both conditions enabled investigation of the specific effects of LLM guidance on user behavior, providing a foundation for further questions regarding trust and reliance on LLM-based guidance. Study 2 examined how interaction with erroneous LLM-based guidance influences trust, mental workload, reliance and gaze behavior. For clarity, the methodology is presented separately for Study 1 and Study 2.

3.2 Study 1: High Accuracy of LLM-Based Guidance

3.2.1 Study Design

RQ2 is addressed using data from a user study conducted in August 2025 as part of a broader experimental project (Madritsch et al., 2026), hereafter referred to as Study 1. Study 1 employed a mixed-methods design comparing an AR-based learning system with a PDF display (*AR+PDF*) and the same system augmented with an integrated LLM-based assistant (*AR+LLM*) under highly accurate LLM-based guidance. LLM-based guidance was manipulated as a between-subjects factor, while task difficulty served as a within-subjects factor operationalized through three sequential juice-mixing tasks of increasing complexity (tasks A–C), completed in a fixed order. In the *AR+PDF* condition, participants could use a PDF manual embedded in the AR interface, whereas the *AR+LLM* condition additionally included an LLM-powered assistant. The study examined subjective (trust in AI, mental workload) and objective (gaze duration) measures. The aim of Study 1 was to compare AR interaction with and without LLM-based guidance and to establish a baseline for understanding how the presence of LLM-based guidance influences user interaction in AR environments.

3.2.2 Participants

Participants for Study 1 were recruited through mailing lists and social media posts during the first two weeks of August 2025. Each participant received a €20 gift card compensation. The sample in the dataset comprises of 36 participants, aged between 18 and 55 years (M = 28.94, SD = 8.14; 14 female, 21 male, 1 non-binary). Based on age, gender, AR experience, and AI familiarity, participants were assigned to two independent groups. Participants assigned to the *AR+PDF* group (7 female, 11 male) were aged 18–55 years (M = 29.28, SD = 10.08), with AR experience ranging from *None* to *Regular use (*M = 1.00*,* SD = 0.91) and AI familiarity ranging from *Slightly familiar* to *Very familiar* (M = 2.67, SD = 0.97). Participants assigned to the *AR+LLM* group (7 female, 10 male, 1 non-binary) were aged 18–41 years (M = 28.61*,* SD = 5.85), with AR experience ranging from *None* to *Regular use* (M = 0.94, SD = 0.94) and AI familiarity ranging from *Slightly* to *Very familiar* (M = 2.44, SD = 0.78).

3.2.3 Experimental Platform

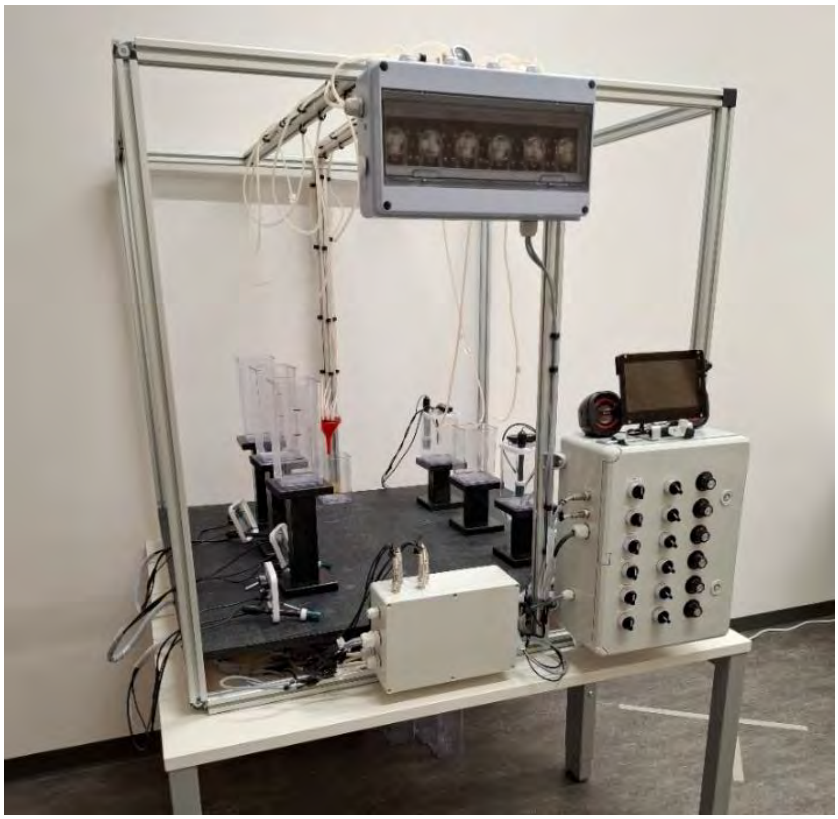
The interactive system used in this study was originally developed by Madritsch et al. (2026) as part of research project *Data-Driven Immersive Analytics in Digital Industries* at Know

Center Research GmbH (Know Center Research GmbH, 2026). The present thesis did not involve the development of the AR system or task design. Instead, an existing implementation was used as the experimental platform for the conducted user studies. A detailed description of the system architecture, hardware setup, and software implementation is provided in the original publication (Madritsch et al., 2026). A brief overview relevant to this thesis is provided below.

Study 1 was conducted using the Juice Mixer testbed (Figure 3), a laboratory installation developed to simulate an industrial machine environment (ElSayed et al., 2025). The setup consisted of six vessels mounted on corresponding stands, equipped with temperature and pH sensors. Pump operation was controlled through a pump control panel with buttons for activation and flow rate adjustment. A result vessel was located in the central rear area of the testbed.

Figure 3

The Experimental Juice Mixer Testbed Used in the Study



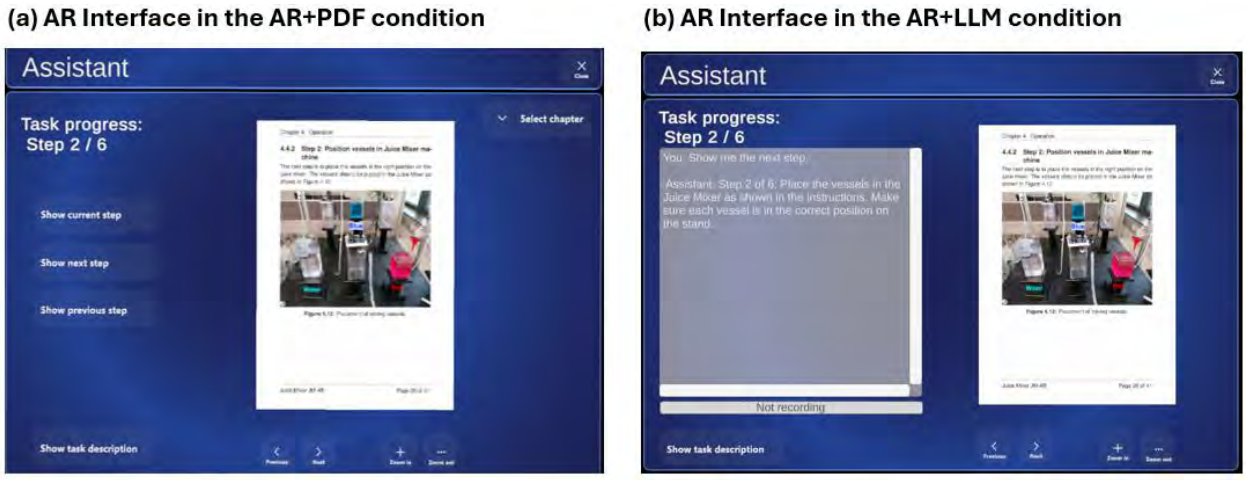
Note. The Juice Mixer testbed, including vessel stands on the left and right side of the testbed, and pump control unit.

3.2.3.1 Technical Implementation

AR System and Eye Tracking Set. The AR application was developed in Unity3D[5] using the Mixed Reality Toolkit 3 (MRTK3)[6] and deployed on a Microsoft HoloLens 2[7] device. The application was streamed to the headset using the Holographic Remoting Player[8]. It integrated three main components: (1) a document-based interface providing procedural instructions in PDF, (2) an LLM-based conversational assistant, as presented in Figure 4, and (3) spatial AR guidance aligned with the physical task environment. Visual guidance was implemented through wireframe cube markers (AR cues) positioned within the physical environment to highlight relevant components. The markers were continuously visible and changed color when fixated on to indicate the currently relevant component. Eye-tracking functionality of the HoloLens 2 was used to capture participants' gaze behavior during task execution.

Figure 4

AR Interface Conditions



Note. The figure illustrates the user's point of view of the AR interfaces displayed through the AR headset. Panel (a) shows the AR interface in the *AR+PDF* condition, comprising task documentation in PDF format; panel (b) shows the AR interface in the *AR+LLM* condition with LLM-based guidance, comprising conversational AI assistant chat window alongside the PDF task documentation.

[5] https://unity.com/de/releases/editor/whats-new/2022.3.33f1
[6] https://learn.microsoft.com/en-us/windows/mixed-reality/mrtk-unity/mrtk3-overview/
[7] https://learn.microsoft.com/de-de/hololens/
[8] https://learn.microsoft.com/de-de/windows/mixed-reality/develop/native/holographic-remoting-player

Conversational AI Assistant and LLM Implementation. The LLM-based assistant was implemented using the OpenAI Assistant API[9] and followed an in-context learning (ICL) approach (Duricic et al., 2024; Madritsch et al., 2026). User input was captured via a Bluetooth microphone, transcribed using OpenAI Whisper[10], processed by the LLM, and returned as both text within the AR interface and synthesized speech using OpenAI TTS[11]. The assistant generated structured outputs that were interpreted by the AR application to display instructions and trigger AR cues within the physical environment. The assistant generated its responses based on a structured representation of the instruction manual, which was incorporated into the prompt context. To enable this integration, the manual was converted into a structured JSONL-based dataset and incorporated into the assistant's knowledge base using ICL. This ensured that the assistant's responses were grounded in the same task-specific documentation available to participants through the PDF interface.

3.2.3.2 Instructional Materials

The study utilized a structured instruction manual designed to guide participants through a multi-step juice-mixing procedure. The manual was provided as a PDF within the AR interface and contained step-by-step instructions for all tasks. Each task was organized into six main procedural steps, further divided into multiple substeps, and supported by illustrative images and cross-references to relevant sections of the manual to facilitate navigation and task execution. Participants could access and navigate the manual at any time during task performance using interface controls such as zoom and page navigation. Both instructional modalities, the PDF manual and the LLM-based assistant, were based on a shared underlying source of information, differing only in their mode of interaction and presentation.

For clarity, the terms *LLM-based assistant*, *LLM assistant, conversational AI assistant*, and *the assistant* are used interchangeably throughout this thesis to refer to the LLM-based system participants interacted with during the experiments. Similarly, the terms *LLM-based guidance* and *LLM outputs* are used interchangeably to refer to the responses generated by the LLM-based assistant during task interaction. The term *AR with AI* is used when referring more generally to the integrated conversational assistance functionality within the AR environment.

[9] https://platform.openai.com/
[10] https://openai.com/index/whisper/
[11] https://developers.openai.com/api/docs/guides/text-to-speech

3.2.4 Study Procedure

Data collection took place between August and September 2025 at Know Center Research GmbH in Graz, Austria. Each session lasted up to 90 minutes and consisted of three phases: briefing and training, the study phase, and debriefing.

At the beginning of the session, participants were introduced to the study procedure and the laboratory equipment. After providing informed consent, participants completed a training task designed to familiarize themselves with the Juice Mixer testbed, the AR interface and the available instructional tools. Depending on the assigned conditions, participants used either a PDF instruction manual or a combination of the PDF manual and the LLM-based assistant; they were shown how to navigate the task steps and use the available support features within the AR environment and then completed demographic (see Appendix A.1) and Trust in AI questionnaires. In the study phase, participants completed three juice-mixing tasks of increasing complexity in a fixed order, following instructions available via the PDF interface and/or LLM-based guidance. After each task, participants completed Trust in AI questionnaires and the Raw NASA Task Load Index. In the debriefing phase, participants were given the opportunity to provide feedback on their experience, were debriefed and thanked for their participation.

3.2.5 Task Design

The study adopted the task structure introduced by Madritsch et al. (2026), comprising three tasks of increasing complexity, completed in a fixed order (A–C). In each task, participants performed a standardized multi-step procedure using the AR system with the PDF manual and/or LLM-based assistant.

Task A. Participants followed a structured recipe involving two mixing vessels and three liquids (plain water, blue-colored water, and red-colored water).

Task B. Task complexity increased as participants were required to navigate the instructions in a non-linear manner: in addition to the standard procedure, participants had to perform safety checks, inspect sensor readings, and identify a malfunctioning pump. This required consulting different parts of the instructions rather than following a strictly sequential workflow.

Task C. The physical testbed's initial setup differed from the setup shown in the instructions. Participants had to identify and correct the mismatch before continuing the procedure.

3.2.6 Measures

3.2.6.1 Subjective Measures

Mental Workload. Mental workload was assessed after each task using the Raw NASA-TLX (R-TLX), the unweighted version of the NASA-TLX Index (Hart & Staveland, 1988; see Appendix A.2). The scale captures workload across six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration, using a 6-item, 20-point scale (1–20). To ensure consistency across dimensions, the Performance subscale was reverse-coded, as higher original scores indicate better performance (i.e., lower workload). An overall mental workload score was computed as the mean of all six subscales, with higher values indicating higher mental workload (hereafter referred to also as MWL). The questionnaire was administered after each of the three tasks, allowing changes in mental workload to be examined across tasks and conditions.

Trust in AI. Trust in AI was assessed using an adapted version of the Trust in XAI Index, originally proposed by Hoffman et al. (2023) and later validated by Perrig et al. (2023). The scale captures user's perceived reliability and confidence in systems that provide automated recommendations during task execution. In line with prior validation work, Item 6 was excluded due to its negative impact on internal consistency and model fit; Item 7 was removed, as it was not applicable to the AR-based experimental setting. Minor wording adjustments were introduced to ensure suitability across measurement points (baseline, i.e., before task execution, and post-task), modifying verb tense and adapting phrasing to the experimental condition. The resulting instrument comprised six items rated on a 5-point Likert scale (1 = *Strongly disagree* to 5 = *Strongly agree*) (see Appendix A.3–A.6). An overall score was calculated as the mean across all items, with higher values indicating greater trust in AI. The questionnaire was administered after the training phase/before task exposure to assess baseline values and after each of the three tasks, allowing changes in trust to be examined over time/across tasks and conditions.

Although the questionnaire was adapted from an existing Trust in XAI scale, its items were considered suitable for assessing trust in both conditions (with LLM-based and with automated system guidance) within the experimental context. For consistency across studies and to maintain terminological clarity in relation to Study 2, the measure is referred to throughout the thesis as the *Trust in AI* questionnaire.

3.2.6.2 Objective Measures

Eye gaze behavior. Gaze-based visual attention allocation was measured using eye tracking data recorded by the Microsoft HoloLens 2 AR headset, which supports validated eye tracking and Areas of Interest (AOI) mapping. Within the AR interface, two AOIs were defined: the PDF display and the conversational AI assistant's chat window. Additional AOIs were defined for other interface elements; however, the present analysis focuses on the PDF and chat window interface. Gaze events within these AOIs were recorded with millisecond precision. Fixations were identified based on continuous gaze within a defined AOI, and total gaze duration was computed by summing fixation times, calculated as the difference between the start and end timestamps of each gaze event. These measures served as indicators of how participants allocated their visual attention between static instructional content (PDF manual) and the LLM-based assistant during task execution.

3.2.7 Data Analysis Strategy

The dataset comprises responses from 36 participants (see Section 3.2.2). Descriptive statistics, including means, medians, and standard deviations, were calculated for all variables to summarize central tendencies and distributional patterns. These summaries provided an initial overview of the data and informed subsequent analyses. All measures were examined across three tasks and both experimental conditions (*AR+PDF* and *AR+LLM*) to capture patterns in participants' responses prior to conducting further statistical analyses.

The internal consistency of multi-item scales was assessed using Cronbach's alpha. Reliability coefficients were interpreted according to established thresholds ($\alpha \geq .70$ acceptable, $\alpha \geq .80$ good, $\alpha \geq .90$ excellent).

To address Research Question 2, correlation analyses were conducted as an initial exploratory step to examine relationships between variables. Depending on the distribution of data, either Pearson's correlation coefficient (for approximately normally distributed variables) or Spearman's rank correlation coefficient (when normality assumptions were not met) was applied. Statistical significance was evaluated at a threshold of $p < .05$. In addition, data visualization techniques, including boxplots and scatterplots, were used to illustrate patterns in trust, subjective workload, and gaze duration across tasks and experimental conditions.

3.2.8 Data Preparation

The data were fully anonymized prior to analysis, with all personally identifiable information removed[12]. Data was analyzed using statistical software RStudio (version 4.5.1.). Prior to analysis, the data were inspected for missing values, inconsistencies, and outliers. Variables were formatted and organized to ensure compatibility with the planned statistical procedures.

3.3 Study 2: Reduced Accuracy of LLM-Based Guidance

3.3.1 Study Design

Study 2 was conducted between December 2025 and January 2026 as a within-subject experiment in a controlled laboratory environment[13]. It builds on the design of Study 1 by retaining the same system setup, task structure, and interaction modalities from the *AR+LLM* condition to ensure comparability across studies. It extends this approach by introducing controlled errors into the LLM-based guidance during task A. Tasks B and C were performed with fully accurate LLM-based guidance (100% accuracy), allowing observation of user behavior following exposure to imperfect AI-generated guidance early in the interaction.

Participants completed three consecutive tasks of increasing complexity in a fixed order (tasks A–C) while interacting with three sources of information: an AR interface with (1) an LLM-based assistant providing procedural guidance upon the user's request, and (2) a fully accurate PDF manual, and (3) a physical juice-mixing testbed augmented with AR visualizations. Each task required participants to follow a multi-step procedure, including preparing and connecting components, operating the system, and completing the mixing process.

Errors introduced in task A were intentionally designed to appear plausible within the task context rather than being obviously incorrect. The erroneous instructions could only be identified when participants critically evaluated the instructions against the task context or cross-checked them with the PDF documentation. A fixed experimental design was used to preserve the temporal structure of interaction, as Study 2 examined how early errors in LLM-based guidance influenced trust and behavior over time. Counterbalancing task order would have reduced the ability to

[12] All raw data, related to this study, are openly available on Zenodo: https://doi.org/10.5281/zenodo.20389976
[13] Findings from Study 2 were accepted as a short paper titled "*An Investigation of Trust and Overreliance on Conversational Instructions in Augmented Reality"* to the ACM Conference on Conversational User Interfaces (CUI 2026) at the time of thesis submission.

investigate these sequential effects. Consequently, learning effects across tasks cannot be fully excluded.

Subjective (Trust in AI, Trust in LLM, mental workload), objective (gaze-based attention allocation) and behavioral (task performance, overreliance behavior) measures that were collected during the study to assess how variations in LLM-based guidance accuracy influenced user interaction with the system are described in Section 3.2.6.

Participants were deliberately not informed about the manipulation of the LLM-based guidance, ensuring that their responses reflected natural trust formation and reliance behavior. This enabled the examination of trust adjustment following early system errors, potential recovery of trust after subsequent accurate performance, patterns of behavioral reliance under imperfect LLM-based guidance, and the relationship between trust, mental workload, and gaze-based attention allocation.

3.3.2 Participants

Participants for Study 2 were recruited through mailing lists and social media posts between November and December 2025. To ensure suitability for the experimental tasks, potential participants completed a short pre-screening questionnaire assessing familiarity with the Juice Mixer setup. Inclusion criteria required participants to be between 18 and 65 years of age, have at least intermediate English proficiency, and no prior experience with the Juice Mixer testbed or related studies. Participation was voluntary and all participants provided informed consent. Each participant received a €20 gift card compensation.

The final sample consisted of 18 participants (7 female, 11 male) aged between 20 and 42 years (M = 29.72, SD = 5.77). All participants reported sufficient English proficiency, with 16 indicating advanced proficiency and two reporting native or bilingual proficiency. Participants reported relatively high familiarity with AI systems, measured on a 5-point Likert-type scale ranging from *Not at all familiar* to *Extremely familiar* (M = 3.89, SD = 0.83), and limited familiarity with AR (M = 1.50, SD = 0.71). Comfortability with AI and AR systems was assessed using a 5-point Likert scale ranging from 1 (*Not comfortable at all*) to 5 (*Very comfortable*). Participants reported relatively high comfort with AI systems (M = 4.00, SD = 1.03) and more moderate comfortability with AR (M = 3.17, SD = 1.29). Prior experience with AR assistants, measured on a 4-point Likert-type scale (*None/Tried once or twice/Occasional use/*R*egular use),*

was minimal (M = 1.06, SD = 0.24). The sample size of 18 participants was chosen to enable comparison with the results of Study 1.

3.3.3 Experimental Platform

Study 2 built directly on the system used in the *AR+LLM* condition of Study 1 and was deployed on a Microsoft HoloLens 2 device. We use the same apparatus, interface configuration, eye-tracking setup, and LLM-based assistant as in Study 1 (see Section 3.2.3). The system consisted of three components: (1) a document-based interface providing procedural instructions in PDF format, (2) an LLM-based assistant, and (3) spatial AR guidance aligned with the physical task environment (Figure 5).

Figure 5

User View of the AR Assistance Interface and Task Environment

Note. The figure presents the participant's view of the AR-supported Juice Mixer task environment, including the physical testbed, AR cues (Visual Cue AOI), the LLM-based assistant interaction area (Chat Window AOI), and the PDF-based instruction display (PDF Display AOI).

3.3.3.1 Technical Implementation

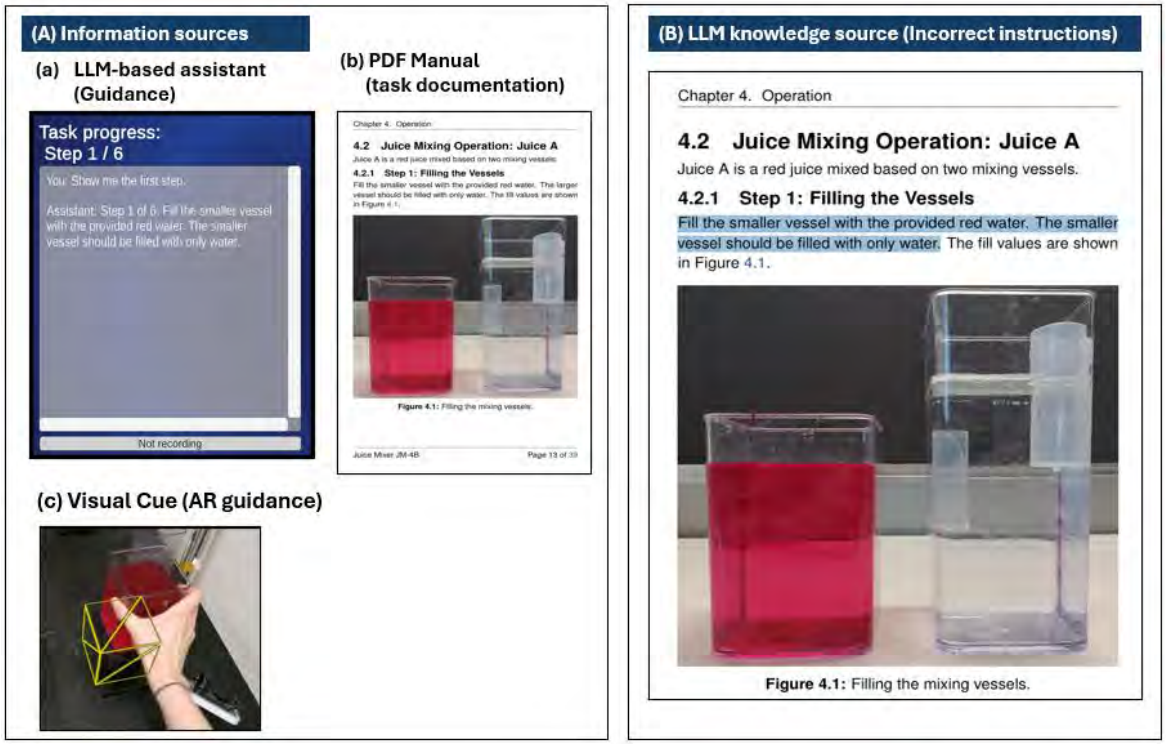
For details about the technical implementation of the AR system, eye tracking, conversational AI and LLM Implementation, see section 3.2.3.1.

3.3.3.2 Instructional Materials

Two versions of the instruction manual were used (Figure 6). The original manual contained correct instructions for all tasks and was available as a PDF within the AR interface. Instructions were presented step-by-step, with six main steps and multiple substeps per task (task A: 16, B: 18, C: 16), supported by illustrative figures and cross references to relevant sections of the manual. Participants could access and navigate the PDF throughout the tasks.

Figure 6

Information Sources and LLM Guidance Manipulation in the AR Task Environment



Note. The figure presents the AR task environment and information sources available to participants. Panel (A) presents the information sources used during the task: (a) the LLM-based assistant interface, (b) task documentation in PDF format, and (c) AR visual cues for AR guidance. Panel (B) shows the modified task documentation used as the knowledge source for the

LLM-based assistant. When participants interacted with the LLM-based assistant, the response was also displayed on the corresponding page on the PDF interface.

A second, structured version of the manual was used as the knowledge source for the LLM-based guidance for task A, where we intentionally introduced errors into selected substeps (see Appendix C). The errors introduced in task A were designed to appear plausible within the task context rather than obviously incorrect. Participants could identify the errors either by carefully assessing the physical setup or by comparing the LLM-based guidance with the PDF instructions. In addition, the relevant PDF page, including the corresponding figure, was displayed next to the chat window, making it possible to visually check whether the guidance matched the instructions. In tasks B and C, the assistant used the unmodified manual and produced fully accurate responses. The PDF manual remained accurate across all tasks. Participants could consult both information sources and cross-check information throughout task execution.

3.3.4 Study Procedure

Data collection took place between December 2025 and January 2026 at Know Center Research GmbH in Graz, Austria. Each session lasted up to 90 minutes and consisted of three phases: briefing and training, the study phase, and debriefing.

In the briefing and training phase, participants were introduced to the study. After providing informed consent, participants completed a guided training session to familiarize themselves with the Juice Mixer testbed, the Microsoft HoloLens 2, and the AR interface. They learned how to use the PDF instruction manual and interact with the conversational AI assistant. This step ensured that all participants understood the basic interface functions before beginning the experimental tasks. They then completed questionnaires to assess trust in AI and trust in LLMs before interacting with the system (to assess baseline values), along with a demographic questionnaire. In the study phase, participants completed three tasks in a fixed order (A–C) while wearing the AR headset and interacting with the LLM-based assistant and Juice Mixer testbed. During this phase, gaze data, task performance, and indicators of overreliance were recorded. After each task, participants completed the R-TLX and trust questionnaires. At the end of the session, participants were informed about the manipulation of the accuracy of the LLM-based guidance in task A and were given the opportunity to ask questions and provide feedback on their experience.

3.3.5 Task Design

The study extends the study design of Study 1 (see Section 3.2.5) by introducing controlled errors in LLM output. Task execution was supported by an LLM-based assistant and a PDF manual within the AR interface.

Task A. Participants followed a sequential procedure using two mixing vessels and three liquids. Task A included an experimental manipulation in which the LLM-based assistant provided incorrect instructions in three out of 16 substeps (see Section 3.3.3.2), resulting in approximately 81.25% LLM-based guidance accuracy.

Task B. Task complexity was increased by introducing a troubleshooting component that extended the standard procedure, i.e., participants were required to consult additional chapters and figures within the documentation to obtain more detailed instructions and complete a task step successfully. The LLM-based assistant operated with 100% guidance accuracy.

Task C. Participants were required to assess and correct a pre-configured system setup before continuing the task, as several items were already positioned on the testbed at the start of the task, but their configuration differed from the initial setup described in the instructions. This mismatch was part of the physical task setup and was independent of the LLM-based guidance, which operated with 100% guidance accuracy throughout the task. Participants therefore had to identify and correct the physical setup mismatch before proceeding, increasing task complexity.

3.3.6 Measures

3.3.6.1 Subjective Measures

Mental Workload. Mental workload was assessed using the Raw NASA-TLX (Hart & Staveland, 1988; Appendix B.2) and the same approach as in Study 1 (see Section 3.2.6.1).

An overall workload score was computed as the mean of all six subscales, with higher values indicating higher mental workload. The questionnaire was administered after each of the three tasks, allowing changes in workload to be examined across tasks and conditions.

Trust in AI. Trust in AI was measured using the same adapted version of the Trust in XAI Index (with minor wording adjustments; see Appendix B.3–B.4) as described in Study 1 (for detailed description, see Section 3.2.6.1). An overall score was calculated as the mean across all items, with higher values indicating greater trust. The questionnaire was administered after the training phase/before task exposure to assess baseline values and after each of the three tasks, allowing changes in trust to be examined across tasks.

Trust in LLM. Trust in large language models (LLMs) was assessed using the Trust-in-LLMs Index (TILLMI; De Duro et al., 2025), which captures trust in conversational AI systems by distinguishing between cognitive (perceived competence and reliability) and affective trust (emotional response to interaction). This distinction is particularly relevant in the present study, where participants interacted with an LLM-based assistant in an AR environment and where trust may be influenced both by system performance (e.g., accuracy of instructions) and by interaction qualities (e.g., conversational guidance). The TILLMI was administered at baseline (prior to the task exposure) and after each task to assess changes in trust across tasks (see Appendix B.5–B.6). Trust in LLM scores were calculated as the means across all items for each task, with higher values indicating greater trust in the LLM.

3.3.6.2 Objective Measures

Eye Gaze Behavior. Gaze-based visual attention allocation in Study 2 was measured using the same eye-tracking setup and AOI definitions as in Study 1 (see Section 3.2.6.2). As in Study 1, gaze data were recorded using the Microsoft HoloLens 2 AR headset; the chat window and the PDF display were defined as primary AOIs; gaze events within these AOIs were recorded with millisecond precision, and fixation durations were calculated based on the time between the start and end of each gaze event. For each participant and task, total gaze duration time was computed by summing fixation durations separately for each AOI.

Building on the measurement approach used in Study 1, Study 2 extends the analysis by explicitly linking gaze allocation to reliance behavior. The two AOIs represent distinct information sources: the chat window provides AI-generated guidance, whereas the PDF display allows access to the original instructions for verification and independent information seeking. Accordingly, increased attention to the chat window was interpreted as reliance on LLM-based guidance, while greater attention to the PDF display indicated verification behavior.

3.3.6.3 Behavioral Indicators

Task Performance. Task performance was operationalized for each participant as the proportion of correctly completed substeps within each task (task A: 16; B: 18; C: 16). Each substep was coded by the experimenter as correct (1) or incorrect (0). Performance scores were computed as the ratio of correct to total substeps (Madritsch et al., 2026).

Behavioral Overreliance. Behavioral overreliance was operationalized based on prior work on automation misuse, defined as the tendency to accept incorrect system recommendations without sufficient verification, often resulting in commission errors (Parasuraman & Riley, 1997; Skitka et al., 2000). Following Buçinca et al. (2021, p. 188:9), overreliance was quantified as the "percentage of agreement with the AI when the AI made incorrect predictions." In task A, where erroneous AI-generated responses were experimentally introduced, overreliance was calculated as

$$\text{Overreliance Rate} = \frac{\text{Number of commission errors}}{\text{Total number of AI errors presented}}$$

where commission errors refer to instances in which participants followed incorrect LLM-based guidance. A trial was coded as overreliance when participants followed incorrect LLM-based guidance without verification or correction, or when they followed incorrect LLM guidance despite having verified it against the PDF manual. Trials in which participants corrected the error before acting were not classified as overreliance. The overreliance rate was calculated per participant for task A.

3.3.7 Data Analysis Strategy

Data were analyzed using a combination of descriptive and inferential statistical methods to examine the influence of LLM-based guidance accuracy on trust, mental workload, task performance, gaze behavior, and reliance behavior. Participants were categorized based on their behavioral responses to errors in LLM outputs in task A. The *Error* group included participants who committed at least one out of three possible commission errors, interpreted as behavioral overreliance on the AI system. The *No Error* group consisted of participants who did not follow incorrect LLM outputs.

Descriptive statistics (means, standard deviations, medians) were computed for all dependent variables per participant and task to provide an overview of distributions and trends across tasks. To examine the relationship between trust and overreliance, Spearman rank-order correlations were conducted. This non-parametric method was chosen because the overreliance variable represents a count-based measure with a limited range, and normality assumptions were not fully met. Poisson regression analyses were performed to further investigate whether trust predicted the extent of overreliance. In these models, overreliance frequency was treated as the

dependent variable, while trust (in AI and in LLM) measured after task A was used as the predictor, as it reflects participants' evaluation of the system following exposure to its outputs and is therefore more directly linked to reliance behavior. Results are reported as incidence rate ratios (IRRs) to provide an interpretable estimate of how changes in trust relate to the likelihood of following incorrect LLM outputs. However, given the small sample size and exploratory nature of the study, multivariable regression results should be interpreted cautiously. Overdispersion diagnostics were computed to assess the suitability of the Poisson regression models. The ratio of the Pearson chi-square statistic to the residual degrees of freedom (χ^2/df) was calculated for each model. Predictors were analyzed in two main blocks: psychological variables (Trust in AI, Trust in LLM, Raw-TLX, AI comfortability, AR comfortability) and experience variables (AI familiarity, AR experience). We also analyzed Trust in AI and Trust in LLM as predictors.

Overreliance was measured at predefined error opportunities in task A, ensuring that the measure captured responses to controlled LLM outputs. However, due to the limited range of the dependent variable and sample size, results should be interpreted with caution.

For repeated-measures analyses, assumptions of normality and sphericity were evaluated using Shapiro–Wilk tests and Mauchly's test, respectively. Where the assumption of sphericity was violated, Greenhouse–Geisser corrections were applied. Bonferroni-adjusted post-hoc comparisons were conducted where appropriate.

To address RQ3.3, Trust in AI scores after task A from Study 1 were compared to Trust in AI scores after task A from Study 2 using an independent samples t-test. This analysis examined whether reduced accuracy of LLM-based guidance led to lower trust compared to a comparable high-accuracy condition. Baseline trust scores were first assessed to ensure comparability between samples. Assumptions of normality and homogeneity of variance were tested prior to analysis, and Welch's t-test was used where homogeneity of variance was violated.

3.3.8 Data Preparation

Data for Study 2[14] were collected using paper-based questionnaires and eye tracking technology. Responses were anonymized at the point of collection to ensure that no personally identifiable information was recorded. Questionnaire data were subsequently entered into a Microsoft Excel dataset and then converted to CSV format for analysis. Data was analyzed using

[14] All raw data, related to this study, are openly available on Zenodo: https://doi.org/10.5281/zenodo.20389976.

RStudio (version 4.5.1). Prior to analysis, the dataset was inspected for missing values, inconsistencies, and outliers. Variables were formatted and organized to ensure compatibility with the planned statistical procedures. For consistency with Study 1, variable naming and coding schemes (e.g., Likert-scale response), were aligned across datasets where applicable.

3.4 Ethical Considerations

Study 1 received ethical approval from the Ethics Committee of Graz University of Technology (TU Graz) as part of the original research project (Madritsch et al., 2026). Study 2 received separate ethical approval from TU Graz (EK-92/2025), the institution affiliated with the thesis supervisor; the approval document is included in Appendix D.

For both studies, participants provided informed consent prior to participation and were informed that participation was voluntary and withdrawal was possible at any time without consequences. No personally identifiable information was collected, and all data were anonymized prior to analysis. Participants in Study 2 were debriefed after the experiment due to the experimental manipulation of LLM-based guidance accuracy.

4 Results and Interpretation

This section presents the results of the analyses conducted to address the research questions, including reliability analysis, descriptive, correlational, regression, repeated-measures, and between-group analyses, followed by an overview of key findings.

4.1 Study 1

4.1.1 Reliability Analysis

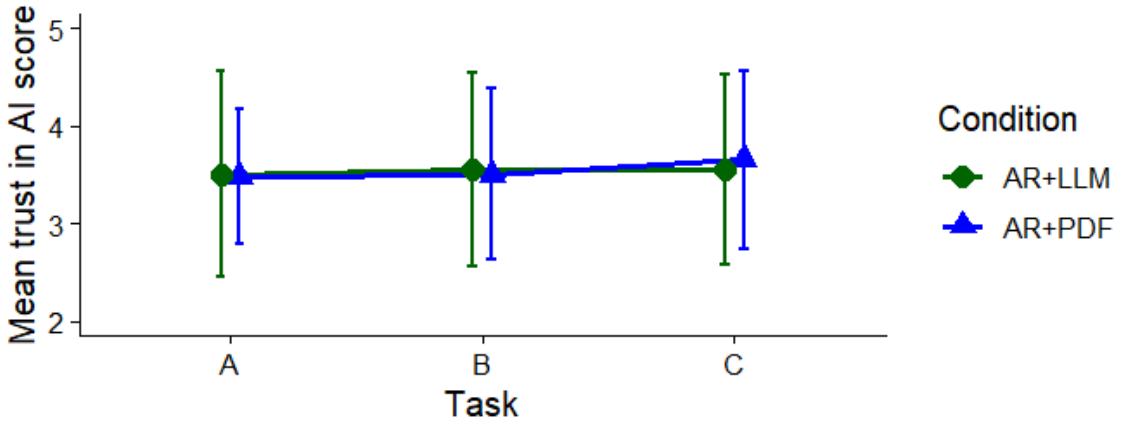
Internal consistency of the questionnaire scales was assessed using Cronbach's alpha for each experimental condition to evaluate scale reliability and consistency across items. Cronbach's alpha values were interpreted using commonly accepted reliability thresholds ($\alpha \geq .70$ acceptable, $\alpha \geq .80$ good, $\alpha \geq .90$ excellent). The *Trust in AI* scale demonstrated excellent internal consistency in both conditions (AI: $\alpha = .97$; AR: $\alpha = .91$). The R-TLX showed high reliability, with excellent reliability in the *AR+LLM* condition ($\alpha = .92$) and good reliability in the *AR+PDF* condition ($\alpha = .84$). These results suggest that the instruments reliably captured the intended constructs.

4.1.2 *RQ2: What Are the Relationships Between Trust in AI, Mental Workload, and Gaze-Based Attention Allocation in AR-Based Learning Environments With and Without LLM-Based Guidance?*

Descriptive statistics were examined to identify overall trends, variability, and potential differences in trust and mental workload across tasks and experimental conditions. Descriptively, participants exhibited moderate levels of trust in AI across both conditions (Figure 7): in the *AR+PDF* condition, trust in AI increased slightly across tasks (task A: $M = 3.49$, $SD = 0.69$; Task B: $M = 3.51$, $SD = 0.88$; Task C: $M = 3.66$, $SD = 0.91$); in the *AR+LLM* condition, trust remained relatively stable (Task A: $M = 3.51$, $SD = 1.05$; Task B: $M = 3.55$, $SD = 0.99$; Task C: $M = 3.56$, $SD = 0.97$), with slightly greater variability than in the *AR+PDF* condition.

Figure 7

Mean Trust in AI Scores Across Tasks and Conditions

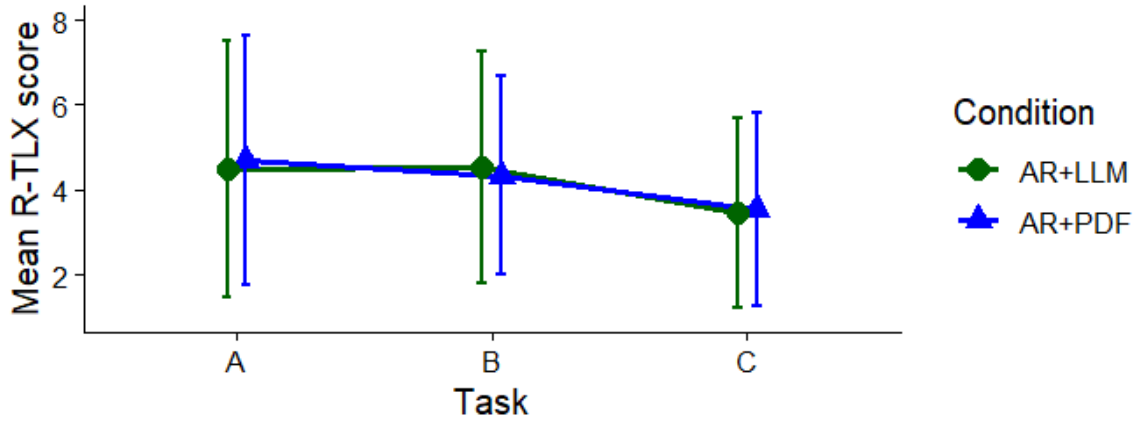


Note. The plot presents mean Trust in AI scores, which were relatively stable across tasks in both *AR+LLM* and *AR+PDF* conditions, with a slight increase observed in task C. Error bars represent ±1 standard deviation.

Mental workload was highest in task A and decreased across subsequent tasks in both conditions (Figure 8): in the *AR+PDF* condition, workload declined from task A ($M = 4.69$, $SD = 2.92$) to B ($M = 4.34$, $SD = 2.32$) and C ($M = 3.55$, $SD = 2.28$); in the *AR+LLM* condition, workload remained relatively stable between tasks A ($M = 4.49$, $SD = 3.03$) and B ($M = 4.54$, $SD = 2.73$), before decreasing in task C ($M = 3.46$, $SD = 2.22$).

Figure 8

Mean Mental Workload Scores Across Tasks and Conditions



Note. The plot presents mean R-TLX (mental workload) scores. They remained relatively stable across tasks A and B and decreased slightly in task C in both conditions (*AR+LLM*, *AR+PDF).* Error bars represent ±1 standard deviation.

Correlation analyses, conducted to investigate whether trust in AI was associated with participants' MWL across tasks and conditions, revealed a consistent negative association between the latter in the *AR+LLM* condition (Table 1), indicating that higher trust was associated with lower MWL across all tasks (rs = −.59 to −.71, ps ≤ .01). In the *AR+PDF* condition similar but weaker negative associations were observed; only the Pearson correlation in task C reached statistical significance (r = −.57, p = .01).

Table 1

Correlations Between Trust in AI and Mental Workload Scores Across Tasks and Conditions

Condition		AR+LLM				AR+PDF			
Task	*n*	*r*	*p*	*ρ*	*p*	*r*	*p*	*ρ*	*p*
A	18	**−.59***	.01	**−.65****	.004	−.33	.18	−.31	.21
B	18	**−.59***	.01	**−.58***	.01	−.38	.12	−.31	.21
C	18	**−.71*****	< .001	**−.64****	.004	**−.57***	.01	−.42	.08

Note. r = Pearson correlation coefficient; ρ = Spearman's rank correlation coefficient. * $p < .05$, ** $p < .01$, *** $p < .001$. Bold indicates statistically significant results.

Gaze behavior varied across tasks and displays. In the *AR+PDF* condition, gaze duration increased from task A to B and then decreased substantially in task C (M = 280,790 → 310,063 → 139,104 ms). A similar pattern was observed in the *AR+LLM* condition, where participants split attention between the PDF and chat window (chat window: M = 223,918 → 253,955 → 176,073 ms; PDF display: M = 150,827 → 180,183 → 99,472 ms) (see Appendix E, Tables E1–E2); however, gaze durations on the chat window were consistently longer than on PDF display.

Correlation analysis examining relationships between gaze duration and mental workload, and gaze duration and trust in AI in the *AR+PDF* condition revealed largely non-significant associations (Table 2). In task A, mental workload showed a moderate positive but non-significant association with PDF gaze duration (r =.43, p =.08; ρ =.42, p = .08), while trust in AI was weakly negatively associated to gaze (r = −.29, p = .24; ρ = −.21, p = .41). In task B, a significant positive association between workload and gaze duration was observed using Spearman's correlation (ρ = .53, p = .03); corresponding Pearson correlation was not significant (r = .39, p = .11). No other significant associations were observed.

Table 2

Spearman and Pearson Correlations between Trust in AI and Gaze Duration, and Between Mental Workload and Gaze Duration on PDF Display Across Tasks in AR+PDF Condition (n=18)

Task	Variables	*R*	*p*	*ρ*	*p*
A	NASA-TLX – PDF gaze	.43	.08	.42	.08
	Trust in AI – PDF gaze	−.29	.24	−.21	.41
B	NASA-TLX – PDF gaze	.39	.12	**.53***	.03
	Trust in AI – PDF gaze	.13	.60	.33	.18
C	NASA-TLX – PDF gaze	−.01	.98	.08	.74
	Trust in AI – PDF gaze	−.003	.99	.15	.56

Note. r = Pearson correlation coefficient; ρ = Spearman's rank correlation coefficient. * $p < .05$, ** $p < .01$, *** $p < .001$. Bold indicates statistically significant results.

Correlation analyses examining relationships between gaze duration on both displays and trust in AI, and gaze duration on both displays and MWL in the *AR+LLM* condition revealed no statistically significant associations (see Appendix E, Table E3). In task A, MWL and PDF gaze showed a moderate positive but non-significant association (r = .41, p = .09; ρ = .35, p = .16),

while all other associations were weak ($|r| \leq .18$). In task B, a similar positive trend was observed ($r = .42$, $p = .08$), whereas relationships involving trust and chat window gaze duration remained small and non-significant ($|r| \leq .33$). In task C, all correlations were weak and non-significant ($|r| \leq .29$, all $p \geq .25$).

4.1.3 Overview of Key Findings

Across both *AR+PDF* and *AR+LLM* conditions, participants demonstrated moderate and relatively stable levels of trust in AI, with slightly greater variability in the *AR+LLM* condition, indicating individual differences in how users perceived automated support. In the *AR+PDF* condition, trust showed a small increase across tasks, whereas it remained stable in the *AR+LLM* condition.

Mental workload in the *AR+LLM* condition increased slightly from task A to B, before decreasing in task C. In the *AR+PDF* condition, MWL decreased across tasks and was significantly lowest in task C, which might indicate increasing familiarity with tasks or a learning effect. In the *AR+LLM* condition, a consistent negative relationship between MWL and trust in AI was observed, indicating that higher trust was associated with lower mental workload across all tasks. In contrast, this association was weaker and only partially significant in task C in the *AR+PDF* condition. This indicates trust and workload became much more strongly associated when LLM-based assistance was present.

Attention allocation, based on gaze duration, differed between conditions. When users were provided with an LLM assistant, their attention was divided between the PDF and the chat window. Within the *AR+LLM* condition, gaze duration was consistently higher on the chat window than on the PDF display. Across both conditions, gaze duration on both displays was highest during task B and declined substantially in task C, possibly reflecting learning effects, task familiarity or reduced mental workload demands across tasks.

Relationships between gaze duration, trust, and MWL were generally weak and inconsistent across both conditions. In the *AR+PDF* condition, only a significant positive association was observed between MWL and PDF gaze duration in task B using Spearman's correlation; all other gaze-related correlations were non-significant. In the *AR+LLM* condition, no statistically significant associations were found between gaze duration, trust in AI, and MWL for either the PDF display or chat window. Overall, gaze-based attention allocation appeared largely independent of trust in AI and only weakly related to cognitive workload across conditions.

4.2 Study 2

4.2.1 Reliability Analysis

This analysis was used to examine the internal consistency of the questionnaire scales and the extent to which items within each instrument reflected the same construct. Cronbach's alpha was calculated for each questionnaire at each measurement point.

The Raw NASA-TLX demonstrated acceptable to good reliability across tasks (α = .78–.89), indicating reliable measurement of mental workload, with slightly higher reliability in later tasks. The Trust in AI scale showed low reliability at baseline (α = .57) but it improved to an acceptable level in task A (α = .80) and reached good reliability in tasks B (α = .89) and C (α = .88). The TILLMI demonstrated questionable to acceptable reliability across measurement points (baseline*:* α = .72; tasks A–C: α = .58–.69), with the lowest reliability observed in task A and a gradual increase in subsequent tasks. Despite these results, the TILLMI was retained because of its theoretical relevance in the present study. Accordingly, results involving TILLMI were interpreted with caution and were considered alongside additional subjective, behavioral, and physiological measures.

4.2.2 Results Addressing Research Questions

4.2.2.1 RQ3.1: How Do Early Errors in LLM-Based Guidance Influence Trust, Reliance Behavior, and Task Performance Across Tasks?

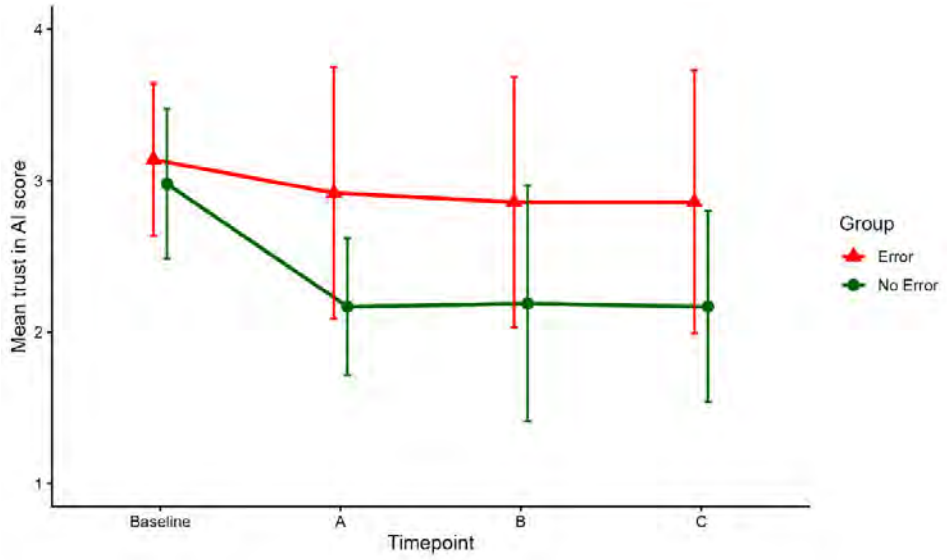
Descriptive statistics, mixed-design ANOVAs and correlation analyses were conducted to examine RQ3.1.

Overreliance and Group Classification. Participants showed moderate overreliance (M = 1.17, SD = 1.20), with 61.11% committing at least one commission error in task A. Error frequency was generally low, with most participants making one error (27.78%), while higher error count (two: 11.11%; three: 22.22%) was less frequent. Error rates descriptively decreased across the three AI error opportunities (from 44% to 33%). Based on these responses, participants were categorized into the *Error* group (n = 11) and the *No Error* group (n = 7).

Trust Development Across Tasks. Trust in AI decreased after task A and remained relatively stable thereafter in both groups (Figure 9). In the *Error* group, trust declined from baseline (M = 3.14) to task A (M = 2.92) and stayed stable across tasks B and C (M = 2.86). In the *No Error* group, it decreased more sharply from baseline (M = 2.98) to task A (M = 2.17) and remained stable thereafter (task B: M = 2.19; task C: M = 2.17) (see Appendix F, Table F1).

Figure 9

Mean Trust in AI Scores Across Timepoints (Baseline, tasks A–C) and Groups (Error, No Error)



Note. The plot presents mean Trust in AI scores across timepoints (baseline and after tasks A–C) for the *Error* and *No Error* groups. Error bars represent ±1 standard deviation. Trust decreased after task A and remained relatively stable thereafter. Across timepoints, participants in the *Error* group reported consistently higher trust in AI than those in the *No Error* group; participants in the *No Error* group showed a larger decline in trust following task A compared to the *Error* group.

A 2 (Group: *Error* vs. *No Error*) × 4 (Task: Baseline, task A–C) mixed-design ANOVA on Trust in AI (Table 3) revealed a significant main effect of Task ($F(3, 48) = 7.15$, $p < .001$, $\eta^2_{(G)} = .10$), indicating changes in trust in AI across tasks. This effect remained significant after Greenhouse–Geisser correction ($p = .003$). While the Group effect approached significance ($p = .07$), neither Group nor the Task × Group interaction reached statistical significance.

Table 3

Mixed ANOVA Results for Trust in AI

Effect	*df*	*F*	*p*	$\eta^2_{(G)}$
Group	1, 16	3.64	.07	.16
Task	3, 48	7.15	**< .001***	.10
Group × Task	3, 48	1.97	.131	.03

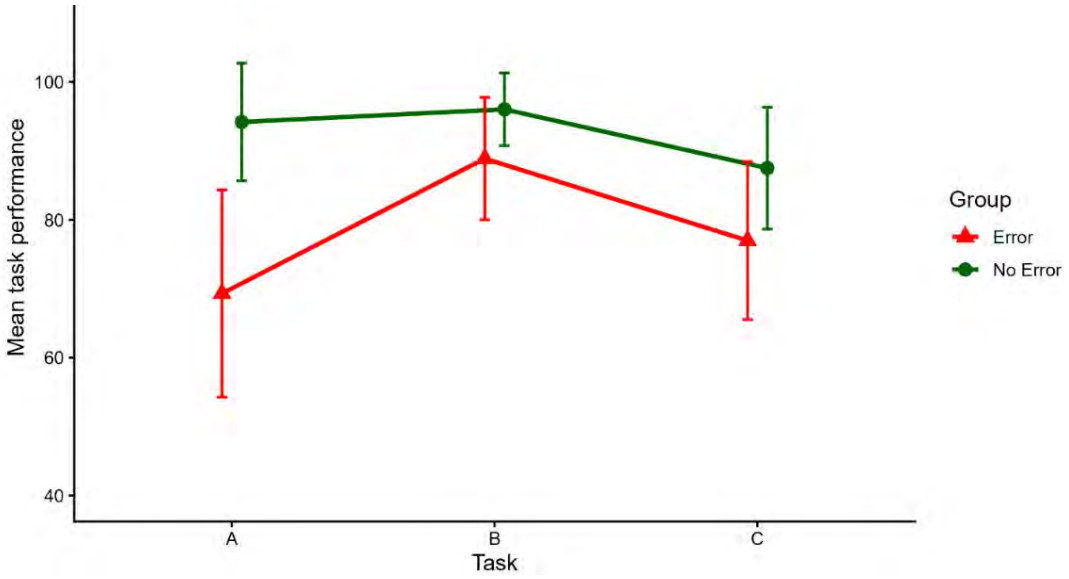
Note. Greenhouse–Geisser corrected p for Task = .003; for interaction = .16. Bold indicates statistically significant results.

Trust in LLM remained relatively stable across tasks in both groups (Error: $M = 2.89–2.95$; No Error: $M = 2.79–3.12$). No significant effects of Task, Group, or their interaction were observed (all ps > .05) (see Appendix F, Tables F2–F3).

Task Performance and Overreliance. Participants in the *No Error* group showed higher performance across tasks A–C (Ms = 94.20, 96.03, 87.50) compared to the *Error* group (task A–C: Ms = 69.32, 88.89, 76.99) (Figure 10). For detailed results, see Appendix F, Table F4.

Figure 10

Mean Task Performance Rates Across Tasks (A–C) And Groups (Error, No Error)

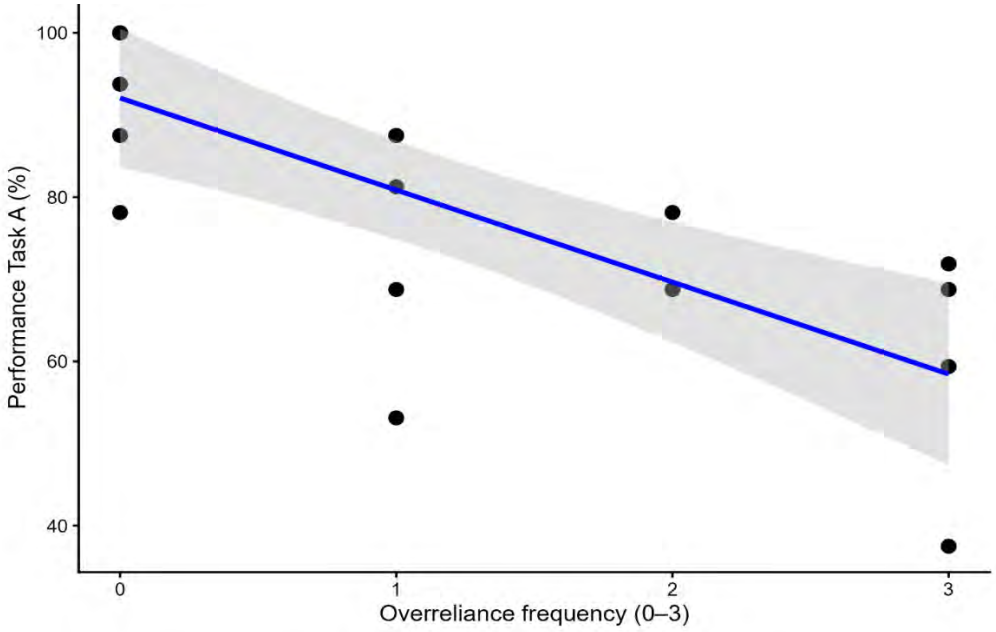


Note. The plot shows mean task performance rates (in %) across tasks A–C for the *Error* and *No Error* groups. Error bars represent ±1 standard deviation. Across all tasks, the *No Error* group demonstrated higher average performance than the *Error* group, whose performance was consistently lower.

A strong negative correlation was found between overreliance frequency and task performance in task A ($r(16) = −.76$, $p < .001$, 95% CI [−.90, −.45]) (Figure 11).

Figure 11

Correlations Between Overreliance Frequency and Performance Rate in Task A



Note. Scatterplot showing the relationship between overreliance frequency and task performance rate (in %) in task A. Each point represents an individual participant (N=18). The solid blue line indicates the linear association between overreliance frequency and performance. The shaded area represents the 95% confidence interval. Higher levels of overreliance frequency were associated with lower task performance scores.

Trust and Overreliance. Spearman correlation analysis further revealed a significant positive association between trust in AI and overreliance in task A (ρ = .59, p = .01), whereas trust in the LLM was not significantly related to overreliance (ρ = .17, p = .51).

A series of Poisson regression models were conducted to examine predictors of overreliance frequency. First, overdispersion diagnostics were computed. They indicated that model assumptions were met (χ^2/df range = 0.82–1.31), supporting the use of Poisson regression. In the trust-only Poisson regression model, trust in AI was a significant positive predictor of overreliance (b = 0.80, p = .010, IRR = 2.22), indicating that each one-unit increase in trust was associated with a 122% increase in expected overreliance. Trust in LLMs was not a significant predictor (p = .642). When psychological variables were included, trust in AI remained a significant predictor (b = 1.19, p = .042; IRR = 3.30), while other variables were not significant. In the full model, including all psychological and experience-related predictors, no predictor

reached statistical significance; trust in AI showed a positive but non-significant association with overreliance (p = .093) (see Appendix F, Tables F5–F6). Overall, trust in AI emerged as the most consistent predictor of overreliance frequency, particularly when examined in isolation or within psychological predictor models.

4.2.2.2 RQ3.2: How Are Trust, Reliance, and Mental Workload Associated With Gaze Behavior and Attention Allocation When Interacting With Erroneous LLM-Based Guidance?

Descriptive statistics, correlation analyses, repeated-measures ANOVAs, and Bonferroni-adjusted pairwise comparisons were conducted to examine RQ3.2.

Across both groups, gaze duration on both displays decreased from task A to task C. For the PDF display, gaze duration declined from M = 333,323 ms to 138,324 ms in the *Error* group and from M = 289,164 ms to 153,852 ms in the *No Error* group. A similar pattern was observed for the chat window (Error: M = 364,600 to 178,620 ms; No Error*:* M = 290,628 to 135,355 ms) (see Appendix F, Table F7). MWL showed a comparable pattern, with R-TLX scores decreasing across tasks in both groups (Error: M = 8.48, 6.39, 6.35; No Error: M = 8.12, 6.00, 5.55 for tasks A–C, respectively). Standard deviations were relatively large, particularly in task A, indicating substantial variability in early task performance (see Appendix F, Table F8).

Correlation analyses showed that higher trust in AI was associated with greater attention to the chat window (r = .33, p = .015; ρ = .29, p = .033) and decreased attention to the PDF display (ρ = −.33, p = .013) (Table 4). This pattern strengthened across tasks, reaching significance in task C for chat window gaze (r = .57, p = .014; ρ = .47, p = .048). No significant associations were found between trust in LLMs and gaze behavior (all ps > .05) (see Appendix F, Table F9).

Table 4

Correlations Between Trust Measures and Gaze Duration (Overall)

Variables	*r*	*P*	*ρ*	*p*
Trust in AI x PDF display gaze	−.24	0.08	**−.33***	0.01
Trust in AI x Chat window gaze	**.33***	0.02	**.29***	0.03
Trust in LLM x PDF display gaze	−.16	0.24	−.19	0.18
Trust in LLM x Chat window gaze	0.12	0.37	0.13	0.34

Note. r = Pearson correlation coefficient; ρ = Spearman's rank correlation coefficient. *p < .05. Bold indicates statistically significant results.

Mental workload was not associated with gaze behavior on either display across tasks (all *p*s > .05). Although it decreased between tasks in both groups, with a significant main effect of Task (F (2, 32) = 6.67, p = .004, $\eta 2(G)$ = .084, corrected p = .007), no significant Group or Group x Task effects were observed (all *p*s > .05; see Appendix F, Tables F10–F11).

Overreliance frequency in task A was not significantly associated with gaze behavior on either display (PDF: $\rho = -.10$, $p = .71$; Chat window: $\rho = -.13$, $p = .60$) in task A. Repeated-measures analysis showed significant main effect of Task for gaze duration on both the PDF display and on Chat window gaze duration (*p*s < .001) (Table 5); no significant Group or Group x Task effects for either display were observed (all *p*s > .05).

Table 5

Mixed-design ANOVA Results for Gaze Duration on PDF display and Chat Window

Display	PDF				Chat Window			
Effect	*df*	*F*	*P*	$\eta^2_{(G)}$	*df*	*F*	*p*	$\eta^2_{(G)}$
Group	1, 16	0.04	0.839	0.001	1, 16	3.27	0.09	0.1
Task	2, 32	**14.45*** **	< .001	0.29	2, 32	**17.68*** **	< .001	0.35
Group × Task	2, 32	0.51	0.606	0.01	2, 32	0.21	0.81	0.01

Note: $\eta^2_{(G)}$ = generalized eta squared. * $p < .05$, ** $p < .01$, *** $p < .001$. Bold indicates statistically significant results.

Bonferroni-adjusted pairwise comparisons further revealed significant differences between all task pairs for both displays (all adjusted *p*s < .05; Table 6). For both the PDF display and chat window, gaze duration progressively decreased from task A to C.

Table 6

Bonferroni-Adjusted Pairwise Comparisons for PDF and Chat Window Gaze Duration Across Tasks (Collapsed Across Groups)

Display	*PDF Display*			*Chat Window*		
Comparison	*t(df)*	*p*	*p_adj*	*t(df)*	*p*	*p_adj*
A vs. B	3.32 (17)	.004	**.012***	3.59 (17)	0.002	**.007****
A vs. C	4.79 (17)	< .001	**< .001*****	5.83 (17)	< .001	**< .001*****
B vs. C	4.03 (17)	.001	**.003****	2.83 (17)	0.012	**.035***

Note. p_adj = Bonferroni-adjusted *p*-value. * $p < .05$, ** $p < .01$, *** $p < .001$. Bold indicates statistically significant results.

4.2.2.3 RQ3.3: To What Extent Does Trust in AI Differ Between Users Interacting With Non-Erroneous (Study 1) and Erroneous LLM-Based Guidance (Study 2)?

To investigate RQ3.3, an independent samples Welch's t-test was conducted to compare mean scores of trust in AI after task A between studies with different participant samples (Study 1 with High accuracy of LLM-based guidance condition and Study 2 with Reduced accuracy of LLM-based guidance condition). The analysis examined whether participants interacting with reduced erroneous LLM guidance reported lower trust in the AI after task A compared to participants interacting with highly accurate LLM guidance.

First, baseline trust in AI (reflecting scores before engaging in tasks) were compared between studies. No statistically significant baseline difference was observed, although the effect size was moderate (d = –0.61), indicating the groups were broadly comparable before task exposure. Further analysis indicated that trust in AI after task A was significantly lower in Study 2 (M = 2.63) compared to Study 1 (M = 3.51, t = -2.84, p = .008, 95% CI [–1.51, –0.25]). The effect size was large (d = –0.95), suggesting a substantial difference between conditions (Table 7).

Table 7

Inferential Statistics for Differences in Trust in AI Between Study 1 and Study 2

Timepoint	Statistic	*p*	*M* (Study 2)	*M* (Study 1)	95% CI (Mean Difference)	Effect Size
Baseline	t = -1.83	.079	3.08	3.48	[-0.86, 0.05]	d = -0.61, 95% CI [-1.3, 0.03]
Task A	**t = -2.84**	**.008**	**2.63**	**3.51**	**[-1.51, -0.25]**	**d = -0.95, 95% CI [-1.76, -0.3]**

Note. Welch's t-test (t) was used due to unequal variances. M = mean; CI = confidence interval; d = Cohen's d. Bold indicates statistically significant results.

4.2.3 Overview of Key Findings

To answer **RQ 3.1**, the findings suggest that early exposure to erroneous LLM-based guidance was associated with reductions in trust in AI across tasks, particularly among participants who did not follow incorrect LLM outputs, suggesting that identifying erroneous LLM outputs may have contributed to reduced trust in the system; trust in LLM remained relatively stable. Participants who demonstrated greater behavioral overreliance on incorrect AI

recommendations also showed lower task performance and higher levels of trust in AI. These findings indicate that users who placed greater trust in the AI system were more likely to follow incorrect recommendations, which in turn was associated with poorer task outcomes.

To **answer RQ 3.2**, the findings suggest that greater trust in AI was associated with increased visual attention toward the LLM assistant's chat interface and reduced attention toward the PDF display, indicating that participants with higher trust in AI allocated more attention toward LLM-based guidance rather than static instructional material. Conversely, lower trust in AI was associated with greater attention toward the PDF display, potentially reflecting increased verification of LLM outputs. Mental workload, trust in LLMs, and overreliance behavior were not significantly associated with gaze allocation. Although weak trends suggested that higher workload may have been associated with greater attention toward the PDF display, these relationships were not statistically significant. However, they may warrant further investigation in studies with larger sample sizes. Overall, these findings indicate that gaze-based attention allocation may reflect users' trust in AI-supported information sources more strongly than perceived mental workload or behavioral overreliance.

To **answer RQ3.3**, the results suggest that exposure to erroneous LLM-based guidance in the initial task led to a significant decrease in trust in AI. While no statistically significant baseline differences in trust were observed between studies prior to task exposure, participants in Study 2 reported significantly lower trust in AI following interaction with erroneous LLM outputs compared to participants interacting with highly accurate LLM-based guidance in Study 1, providing strong evidence that early AI errors negatively impact trust formation. The large effect size indicates that reduced LLM guidance accuracy had a substantial impact on trust formation in the AI system.

5 Conclusion

The concluding chapter discusses the findings in relation to the research questions, followed by the theoretical and practical implications of the results. It further addresses the study limitations, ethical considerations, and directions for future work.

5.1 Discussion by Research Question

5.1.1 RQ1: Factors Determining Reliance on AI

Research Question 1 aimed to identify and structure the factors influencing reliance on AI-generated guidance. Based on literature synthesis, we developed an integrative framework of AI reliance behavior (see Figure 1 in Section 1), conceptualizing reliance as a dynamic behavioral outcome shaped by multiple interacting influences. Within this framework, trust in AI represents an important attitudinal foundation that influences users' willingness to rely on AI-generated outputs. Consistent with theoretical perspectives distinguishing trust from behavioral reliance (Lee & See, 2004), the framework conceptualizes trust as an attitudinal construct that increases the likelihood of reliance (Hoff & Bashir, 2014); hence, reliance partly reflects the behavioral manifestation of trust. This distinction aligns with broader models of attitude–behavior relationships proposed by Ajzen and Fishbein (1975, 1980, as cited in Lee & See, 2004), which suggest that between attitudes and resulting behavior lies a plethora of beliefs, intentions, and contextual constraints that shape whether trust is translated into reliance behavior.

Accordingly, the framework further suggests that the relationship between trust and reliance is indirect and mediated by several mediators, i.e., cognitive and perceptual mechanisms, that help explain why and how trust influences reliance by shaping how users process information and allocate attention. These factors are grounded in theoretical perspectives such as Cognitive Load Theory (Sweller, 1988), Multiple Resource Theory (Wickens, 1984, 2008) and Dual-Process Theories of Decision-Making (Kahneman et al., 1982), which propose that limitations in cognitive resources increase reliance behavior, as individuals become more likely to cognitively offload on- or rely on automated or AI-generated outputs as a way to reduce perceived cognitive strain and minimize effortful cognitive processing. The inclusion of gaze behavior and attention allocation in the framework is further supported by eye-tracking research suggesting that visual monitoring reflects both trust in AI (Ries et al., 2025) and cognitive workload (Ayres et al., 2021).

The framework also highlights that the relationship between trust and reliance is moderated by user characteristics, system properties, model behavior, and task conditions, which help explain when or under what conditions the relationship between trust and reliance changes. In addition, the framework incorporates characteristics specific to LLM-based systems, including conversational fluency, anthropomorphism, and sycophantic behavior, which may increase perceived competence and encourage uncritical acceptance of outputs despite potential inaccuracies (Ibrahim et al., 2025). Beyond individual and system-level influences, broader contextual factors also apply, including organizational context, social norms, and system's interaction design. This is consistent with recent work emphasizing that reliance on AI develops dynamically through interaction, while also being fundamentally shaped by the way such interactions are designed and investigated in the first place: as noted in Eckhardt et al. (2025), before individual or system-related determinants of reliance can be meaningfully interpreted, it is necessary to first consider how study design characteristics, task structure, interaction setup, and experimental characteristics may predispose participants toward particular reliance behaviors and influence how reliance ultimately manifests.

Overall, the proposed framework emphasizes that reliance on AI should be understood as a multidimensional and context-dependent process emerging from the interaction of cognitive, behavioral, technological, and contextual influences. By integrating concepts from trust in automation and AI, cognitive load theory, gaze-based attention research, and recent work on LLM-based interaction, the framework provides a conceptual basis for addressing the research question concerning the factors influencing reliance on AI. Specifically, we synthesize user-, system-, model-, task- and interaction related, and contextual influences that may contribute to appropriate, over- and underreliance behaviors.

5.1.2 RQ2: The Relationships Between Trust in AI, Mental Workload, and Gaze Behavior in AR-Based Learning Environments With and Without LLM-Based Guidance

The findings of Study 1 suggest that relationships between trust in AI, mental workload, and attention allocation differed depending on whether LLM-based guidance was available within the AR environment. Associations between trust in AI and mental workload emerged only in the *AR+LLM* condition, indicating that participants who reported higher trust also tended to experience lower perceived workload during task completion compared to participants in the *AR+PDF* condition. Possible explanation for this is cognitive offloading (Risko & Gilbert, 2016)

where users may have delegated parts of the cognitive processing required for task completion to the LLM assistant. At the same time, the direction of this relationship may work in the opposite direction: participants who experienced lower MWL during interaction with the LLM assistant may have perceived the system as easier to use, more effective, thus lowering their perceived workload, contributing to higher trust evaluations. This is consistent with prior research suggesting that users are more likely to trust AI-supported systems that can reduce mental workload (Shamszare & Choudhury, 2023). Together, these findings support the view that trust and cognitive workload are dynamically interconnected rather than independent constructs. However, mental workload, as measured in the present thesis, is subjective and may have been shaped not only by interaction but also by task difficulty and individual differences in how participants perceived the task demands. This highlights the value of including additional physiological measures of mental workload in future studies, as suggested by Kosch et al. (2023), to better understand how cognitive demands develop during human–AI interaction.

The findings further suggest that the availability of conversational AI guidance changed how participants allocated visual attention across information sources within the AR environment. In the *AR+LLM* condition, where both the LLM assistant and PDF documentation were available, participants spent more time attending to the LLM-based guidance than to the static PDF instructions. This suggests that participants adapted their information-seeking behavior when conversational AI became available, potentially because the interactive guidance provided a more immediate and efficient source of information than static documentation. However, gaze behavior was not strongly associated with trust in AI or mental workload, indicating that visual attention allocation may not directly reflect trust or perceived workload. Instead, gaze patterns may capture engagement with the task, not evaluation of the AI system.

Overall, the findings suggest that AI assistance influences both perceived workload and attention allocation within AR-based learning environments, while gaze-based attention allocation may operate relatively independently from subjective perceptions of trust and cognitive workload.

5.1.3 RQ3.1: Effects of Early Errors in LLM-Based Guidance on Trust, Reliance and Task Performance

Consistent with prior work (Kahr et al., 2024), results showed that early exposure to inaccurate AI guidance influenced trust development, reliance behavior and task performance across tasks. Participants who did not follow incorrect AI recommendations showed reduced trust

in AI in subsequent tasks, suggesting that detecting errors in LLM outputs contributed to trust adjustment during interaction with the system according to its performance. In contrast, participants who followed incorrect LLM outputs appeared to minimally adjust their trust levels following erroneous LLM-based guidance. Furthermore, as in prior work (Chacón et al., 2022; Nourani et al., 2020, as cited in Kahr et al., 2024), trust failed to recover to baseline levels in subsequent tasks even though participants were exposed to fully accurate LLM guidance in subsequent tasks. Accordingly, the findings of Study 2 may reflect the primacy–recency effect, where early interaction experiences disproportionately influence later trust and reliance evaluations (Nourani et al., 2020; Desai et al., 2013, as cited in Kahr et al., 2024).

In addition, participants who demonstrated greater behavioral overreliance on incorrect LLM-based guidance also showed lower task performance and higher levels of trust in AI. These findings indicate that users who placed greater trust in the AI system were more likely to follow incorrect recommendations, which in turn was associated with poorer task outcomes, which is consistent with prior research findings on overreliance on AI (Buçinca et al., 2021).

Additional analyses further identified trust in AI as the most consistent predictor of overreliance frequency as trust in AI remained a significant positive predictor of overreliance even after additional psychological variables were included in the regression models. These findings further support the view that users who place greater trust in AI systems may become more vulnerable to overreliance on erroneous AI-generated recommendations. In contrast, trust in the LLM itself was not significantly associated with overreliance behavior.

5.1.4 RQ3.2: Associations Between Trust, Reliance, and Mental Workload and Attention Allocation During Interaction with Erroneous LLM-Based Guidance

Gaze data provided additional insight into the behavioral mechanisms underlying trust and reliance dynamics. Patterns in gaze allocation suggested that participants with higher trust in AI allocated more attention toward LLM-based guidance and less attention toward the PDF display, potentially reflecting reduced verification behavior. Conversely, lower trust in AI was associated with greater attention toward the PDF display, suggesting continued verification and independent information seeking. These findings indicate that trust may influence not only reliance decisions, but also how users distribute attention across available information sources during AI-assisted interaction as participants who devoted more attention to the LLM-based guidance appeared less likely to look for information in alternative sources. However, prior studies suggest that higher

trust in automation is associated with reduced visual monitoring and shorter gaze durations and vice versa (Hergeth et al., 2016, Ries et al., 2025), while our results showed the opposite pattern: participants with higher trust in AI and stronger overreliance exhibited longer fixation durations on the AI chat window. This discrepancy may be explained by contextual differences in the role of the system: previous research largely examined automation in monitoring or supervisory contexts, where increased gaze reflected scrutiny and distrust, while the current study involved participants learning and performing a novel task in which the AI chat window functioned not as an automated system to supervise, but as an active information source and task aid. In this context, longer fixation durations on the AI interface may not indicate increased monitoring or distrust, but rather increased reliance on the AI as the primary source of guidance. Participants who trusted the AI more may have preferentially allocated their visual attention to LLM-based assistance instead of consulting static instructional materials. Thus, gaze duration likely reflected engagement with and dependence on the AI-provided information, rather than scrutiny of the system itself. This supports the idea that gaze–trust relationships are highly context dependent, and that eye-tracking metrics may capture different underlying processes (Ball and Richardson, 2022), depending on the task and function of the AI system. Moreover, mental workload did not differ between groups and was not significantly associated with gaze allocation as participants who verified and corrected LLM errors did not report higher effort than those who overrelied, suggesting that overreliance is not primarily driven by cognitive demand, but by differences in trust and critical engagement.

Overall, these findings indicate that gaze-based attention allocation may reflect users' trust in AI-supported information sources more strongly than perceived workload or behavioral overreliance.

5.1.5 RQ3.3: Differences in Trust When Interacting with Erroneous and Non-Erroneous LLM-Based Guidance

The comparison between the high LLM-based guidance accuracy and reduced LLM-based guidance accuracy studies further demonstrated that exposure to early erroneous LLM outputs significantly reduced trust following initial interaction. Although no statistically significant baseline differences in trust were observed between the studies prior to task exposure, participants in reduced LLM-based guidance accuracy condition (Study 2) reported significantly lower trust after interacting with erroneous LLM-based guidance compared to participants interacting with highly accurate LLM guidance in Study 1. The observed large effect size further

indicates that reduced accuracy had a substantial negative impact on trust formation in the AI system. Taken together, these findings support prior research suggesting that trust is dynamically shaped through interaction and continuously updated based on how users interpret new evidence and ongoing system performance (Yang et al., 2017, as cited in Kahr et al., 2024). In the present study, participants who identified erroneous outputs appeared to incorporate this new evidence into their evaluation of the AI system, resulting in reduced trust and greater attention to the PDF document. In contrast, participants who failed to detect incorrect guidance appeared less likely to adjust their trust levels or behavior and rather continued to seek and allocate more attention to LLM-based guidance.

Overall, the findings of the present thesis support the view that trust formation in AI systems is highly dynamic and sensitive to early interaction experiences. It also reveals a key mechanism of human–AI interaction failure where trust, unadjusted to erroneous system performance, may reduce information verification efforts, thereby reducing the likelihood of error detection and increasing the risk of overreliance behavior. The findings further extend prior work on trust calibration and the adjustment of trust in response to observed AI system performance (Kahr et al., 2024) to AR environments with integrated AI-generated assistance. However, future research should also examine the contribution of AR-specific environmental factors to trust, reliance behavior, and human–AI interaction dynamics.

5.2 Theoretical Implications

Theoretically, this thesis contributes to cognitive science and human–AI interaction research by integrating key constructs, such as trust as subjective attitude, reliance behavior as its behavioral manifestation, mental workload, and attention allocation, into a unified framework of human–AI interaction, and examining them in the context of conversational AI systems embedded within immersive AR environments, which are relatively underexplored in existing research. In addition, the thesis extends prior findings on trust and reliance behavior in AI-only and screen-based interaction settings as we show that same patterns of trust adjustment and reliance behavior also appear in AR-supported human–AI interaction contexts.

Methodologically, the study contributes by combining subjective self-report measures with objective eye-tracking data and behavioral indicators to capture real-time interaction behavior. The findings also highlight the importance of integrating physiological measures to better understand human behavior and underlying cognitive processes during human–AI interaction.

5.3 Practical Implications

The findings from the present thesis suggest that conversational AI systems integrated into AR environments may encourage users to prioritize AI-generated guidance over independent information seeking and verification, particularly when trust in the system is high. This highlights the need for AI systems and LLM-based tools that are more adaptive and context-aware, particularly in training and decision-support scenarios, and AR systems with integrated AI that could incorporate real-time behavioral and physiological indicators to identify reduced verification behavior and provide adaptive interventions that support appropriate reliance and critical evaluation of AI-generated outputs. Such mechanisms may become particularly important in immersive and decision-critical AR environments, where the conversational nature of LLM interaction and cognitive demands, introduced by AR, may further reduce users' tendency to independently verify information.

In addition, the findings from our studies, that were designed to simulate realistic imperfections of AI systems and observe realistic user behavioral responses, may contribute to the design of AI tools that effectively assist users in different contexts without undermining their (human) judgment, thereby supporting safer and more effective human–AI collaboration.

5.4 Limitations

Sample and Statistical Limitations. Both studies were conducted with relatively small sample sizes, limiting statistical power of results and reducing the likelihood of detecting smaller effects. Especially results of the multivariable regression results should be interpreted cautiously, as these analyses may be underpowered. In addition, several analyses, i.e., analyses examining relationships between gaze behavior, trust, and mental workload, were exploratory, not confirmatory. While they provide useful initial insights, the latter should be interpreted as preliminary and would need further validation in studies with larger samples.

Study Design Limitations. The participant pool was relatively homogeneous, as all participants were recruited from the same city, which may limit the generalizability of the findings to broader and different populations. In addition, tasks were completed in a controlled experimental setting, which does not fully reflect real-world human–AI interaction involving time pressure, longer-term use, or decision-making with real consequences.

Additionally, both studies used a fixed-order task design, meaning all participants completed tasks in the same sequence. This limits the ability to separate experimental effects from

possible learning effects across tasks and limits the chance to investigate how reliance behavior might develop and change over time. However, the design was chosen to preserve the temporal structure of the interaction as the aim of the study was to examine how early timing of AI errors influenced subsequent trust and behavior over time.

In Study 1, the *AR+PDF* condition primarily served as a baseline condition for comparison with the *AR+LLM* condition to examine how the availability of LLM-based guidance influenced user interaction.

In Study 2, errors in LLM outputs were intentionally controlled and systematically introduced to ensure experimental consistency. Although the errors were designed to resemble plausible real-world AI mistakes, they cannot fully capture the dynamic and unpredictable nature of actual AI systems.

Furthermore, while both studies employed AR-based interaction, the present research did not systematically examine the potential influence of characteristics of the AR system, such as level of immersion, interface complexity, or interaction modality. Additionally, the findings are based on a specific AR-supported LLM-based system and may not generalize to other AI systems, interfaces, or application domains.

The studies were also conducted at different times and with different participant samples. The between-study comparisons should therefore be interpreted cautiously.

Moreover, as discussed (see Section 5.1.1), study design characteristics may themselves influence the manifestation of reliance behavior. Therefore, the reliance patterns observed in the present study should be interpreted within the context of the specific experimental setup, task structure, and interaction conditions under which participants engaged with the AI system.

Measurement Limitations. Trust, mental workload, and overreliance were measured during relatively brief interactions across three tasks with limited number of steps. As a result, the studies could not examine sustained reliance behavior or adaptation to repeated errors in AI-generated outputs over longer periods of use.

Participants in Study 2 were classified into *Error* and *No Error* groups based on whether they followed incorrect LLM instructions, which reflects their behavioral response to LLM-based guidance and does not necessarily mean that they were aware of the errors or that they consciously detected them. Similarly, eye-tracking measures captured visual attention allocation to predefined areas of interest, but do not directly reflect trust or underlying cognitive processing. The studies also relied on retrospective self-report measures of trust and mental workload, which

may be influenced by subjective interpretation, response biases, or differences in how participants understood questionnaire items. This may partly explain the lower internal consistency observed for some measures, particularly the baseline Trust in AI and TILLMI scales, potentially reducing measurement precision in some analyses.

Lastly, questionnaires in Study 1 were administered digitally, whereas questionnaires in Study 2 were paper based. Previous research suggests that administration format may influence results, particularly in mental workload assessments (Noyes & Bruneau, 2007, as cited in Babei et al., 2025).

5.5 Ethical Considerations

In designing present thesis, which experimentally examines human responses to erroneous LLMs, several ethical considerations were considered. Prior to data collection, ethical approval was obtained from the relevant Ethics Committee (see Section 3.4). Participation in both studies was voluntary, and participants were informed they could withdraw from the study at any time. All participants provided a signed Informed Consent form. All data was anonymized at the site of collection and handled in accordance with applicable data protection regulations and institutional ethical guidelines. Attention was given to the deceptive element of Study 2, in which participants were intentionally exposed to experimentally controlled erroneous LLM-based guidance. Because revealing the manipulation beforehand could have influenced participants' behavior and compromised the validity of the study, participants were not informed about the erroneous LLM outputs prior to participation. However, all participants were fully debriefed after the study session and informed about the purpose and nature of the experimental manipulation. The study was designed to minimize potential risks to participants, and all tasks were conducted in a controlled, low-risk experimental environment.

5.6 Future Work

Future work might examine the findings of the present thesis under larger and more diverse participant samples to improve statistical power and findings and increase the generalizability of the results across broader populations. It would also be valuable to conduct longitudinal studies to better understand how trust and reliance behavior develop over time and after repeated exposure to AI errors. Such studies could provide further insight into factors that contribute to the persistence or reduction of overreliance behavior. Further work might also

investigate how different task- or AI error-related characteristics influence user behavior in AR-supported AI interactions.

Design-wise, future work might investigate how different AR and AI interface- and interaction patterns influence trust, reliance, verification behavior. This may include different tones of conversational guidance, different modalities of AI-generated guidance, confidence indicators, uncertainty indicators etc. In addition, one might investigate how specific characteristics of AR systems, such as level of immersion, interface complexity, and interaction modality, might shape user behavior.

Finally, future studies should combine behavioral and physiological measures to gain a more comprehensive understanding of user's cognitive processing and behavior during interaction with AI- and AR-supported systems, beyond self-reported experiences alone.

References

Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). Trust in AI: Progress, challenges, and future directions. *Humanities and Social Sciences Communications, 11*(1). https://doi.org/10.1057/s41599-024-04044-8

Anthropic. (2026). *Claude* (Claude Sonnet 4.6) [Large language model]. Retrieved April 10, 2026, from https://claude.ai

Anthropic. (2026). *Claude overview*. Retrieved May April 2, 2026, from https://claude.com/product/overview

Ayres, P., Lee, J. Y., Paas, F., & van Merriënboer, J. J. G. (2021). The validity of physiological measures to identify differences in intrinsic cognitive load. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.702538

Azuma, R. T. (1997). A survey of augmented reality. *Presence: Teleoperators and Virtual Environments, 6*(4), 355–385. https://doi.org/10.1162/pres.1997.6.4.355

Babaei, E., Dingler, T., Tag, B., & Velloso, E. (2025). Should we use the NASA-TLX in HCI? A review of theoretical and methodological issues around mental workload measurement. *International Journal of Human-Computer Studies, 201*, 103515. https://doi.org/10.1016/j.ijhcs.2025.103515

Ball, L. J., & Richardson, B. H. (2022). Eye movement in user experience and human–computer interaction research. In *Neuromethods* (pp. 165–183). Springer. https://doi.org/10.1007/978-1-0716-2391-6_10

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2020). *Does the whole exceed its parts? The effect of AI explanations on complementary team performance* (Version 3). arXiv. https://doi.org/10.48550/arXiv.2006.14779

Becker, E., Wünsche, J., Veith, J. M., Schrader, J., & Bitzenbauer, P. (2025). *From cognitive relief to affective engagement: An empirical comparison of AI chatbots and instructional scaffolding in physics education* (Version 1*).* arXiv. https://doi.org/10.48550/arXiv.2508.06254

Berberette, E., Hutchins, J., & Sadovnik, A. (2024). *Redefining "hallucination" in LLMs: Towards a psychology-informed framework for mitigating misinformation* (Version 1). arXiv. https://doi.org/10.48550/arXiv.2402.01769

Blanco, S. (2025). Human trust in AI: A relationship beyond reliance. *AI Ethics, 5*, 4167–4180. https://doi.org/10.1007/s43681-025-00690-z

Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 454–464. https://doi.org/10.1145/3377325.3377498

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), Article 188, 1–21. https://doi.org/10.1145/3449287

Buchner, J., Buntins, K., & Kerres, M. (2021). A systematic map of research characteristics in studies on augmented reality and cognitive load. *Computers and Education Open, 2*, 100036. https://doi.org/10.1016/j.caeo.2021.100036

Cai, W., Jin, Y., & Chen, L. (2022). Impacts of personal characteristics on user trust in conversational recommender systems. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3491102.3517471

Cao, S., & Huang, C.-M. (2022). Understanding user reliance on AI in assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW2), 1–23. https://doi.org/10.1145/3555572

Chen, V., Liao, Q. V., Vaughan, J. W., & Bansal, G. (2023). *Understanding the role of human intuition on reliance in human-AI decision-making with explanations* (Version 3). arXiv. https://doi.org/10.48550/arXiv.2301.07255

Choudhury, A., & Shamszare, H. (2023). Investigating the impact of user trust on the adoption and use of ChatGPT: Survey analysis. *Journal of Medical Internet Research, 25*, e47184. https://doi.org/10.2196/47184

Cummings, M. L. (2017). Automation bias in intelligent time critical decision support systems. In D. Harris & W.-C. Lu (Eds.), *Decision making in aviation* (pp. 289–294). Routledge.

de Brito Duarte, R., & Campos, J. (2024). Looking for cognitive bias in AI-assisted decision-making. *Proceedings of the Workshops of the Third International Conference on Hybrid Human-Artificial Intelligence (HHAI 2024)*. CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3825/short1-1.pdf

Deck, C., & Jahedi, S. (2015). The effect of cognitive load on economic decision making: A survey and new experiments. *European Economic Review, 78*, 97–119. https://doi.org/10.1016/j.euroecorev.2015.05.004

De Duro, E. S., Veltri, G. A., Golino, H., & Stella, M. (2025). *Measuring and identifying factors of individuals' trust in large language models* (Version 2). arXiv. https://doi.org/10.48550/arXiv.2502.21028

Dietvorst, B. J., & Bharti, S. (2020). People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological Science, 31*(10), 1302–1314. https://doi.org/10.1177/0956797620948841

Duricic, T., Müllner, P., Weidinger, N., ElSayed, N., Kowald, D., & Veas, E. (2024). AI-Powered Immersive Assistance for Interactive Task Execution in Industrial Environments. In *Frontiers in Artificial Intelligence and Applications*. IOS Press. https://doi.org/10.3233/faia241037

Eckhardt, S., Kühl, N., Dolata, M., & Schwabe, G. (2025). A survey of AI reliance. *ACM Computing Surveys, 58*(6), 1–37. https://doi.org/10.1145/3776528

ElSayed, N., Prentner, O., Weidinger, N., & Veas, E. (2025). *The Juice Mixer: A platform for developing and evaluating immersive analytics*. In *2025 IEEE Conference on Human Factors in Immersive Analytics (HFIA)* (pp. 1–5). IEEE. https://doi.org/10.1109/HFIA68651.2025.00005

Fahim, M. A. A., Khan, M. M. H., Jensen, T., Albayram, Y., & Coman, E. (2021). Do integral emotions affect trust? The mediating effect of emotions on trust in the context of human-agent interaction. *Proceedings of the Designing Interactive Systems Conference 2021*, 1492–1503. https://doi.org/10.1145/3461778.3461997

Faruk, L. I. D., Rohan, R., Ninrutsirikun, U., & Pal, D. (2023). University students' acceptance and usage of generative AI (ChatGPT) from a psycho-technical perspective. *Proceedings of the 13th International Conference on Advances in Information Technology*, 1–8. https://doi.org/10.1145/3628454.3629552

Gajos, K. Z., & Mamykina, L. (2022). Do people engage cognitively with AI? Impact of AI assistance on incidental learning. *Proceedings of the 27th International Conference on Intelligent User Interfaces*, 794–806. https://doi.org/10.1145/3490099.3511138

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Gomez, C., Cho, S. M., Ke, S., Huang, C.-M., & Unberath, M. (2025). Human-AI collaboration is not very collaborative yet: A taxonomy of interaction patterns in AI-assisted decision making from a systematic review. *Frontiers in Computer Science, 6*, Article 1521066. https://doi.org/10.3389/fcomp.2024.1521066

Gonnermann-Müller, J., & Krüger, J. M. (2024). Unlocking augmented reality learning design based on evidence from empirical cognitive load studies—A systematic literature review. *Journal of Computer Assisted Learning, 41*(1). https://doi.org/10.1111/jcal.13095

Google DeepMind. (2026). *Gemini*. Retrieved April 2, 2026, from https://deepmind.google/models/gemini/

Gupta, A., Mund, A., Roy, S., Garg, P., & Yadav, D. K. (2025). Trust in AI systems: A social-psychological investigation of human–AI collaboration. *TPM: Testing, Psychometrics, Methodology in Applied Psychology, 32*(S7), 428–446. https://tpmap.org/submission/index.php/tpm/article/view/2133

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

He, G., Kuiper, L., & Gadiraju, U. (2023). Knowing about knowing: An illusion of human competence can hinder appropriate reliance on AI systems. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18. https://doi.org/10.1145/3544548.3581025

Heersmink, R., de Rooij, B., Clavel Vázquez, M. J., & Colombo, M. (2024). A phenomenology and epistemology of large language models: Transparency, trust, and trustworthiness. *Ethics and Information Technology, 26*(3). https://doi.org/10.1007/s10676-024-09777-3

Hergeth, S., Lorenz, L., Vilimek, R., & Krems, J. F. (2016). Keep your scanners peeled. *Human Factors, 58*(3), 509–519. https://doi.org/10.1177/0018720815625744

Himavamshi, S., Bharath, A., Srikanth, G., & Balaji, K. (2024). Exploring the frontiers: A comprehensive review of augmented reality and virtual reality in manufacturing and

industry. *International Journal of Current Science Research and Review, 7*(9). https://doi.org/10.47191/ijcsrr/v7-i9-38

Hoff, K. A., & Bashir, M. (2014). Trust in automation. *Human Factors, 57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science, 5*. https://doi.org/10.3389/fcomp.2023.1096257

Hunter, R., Moulange, R., Bernardi, J., & Stein, M. (2024). *Monitoring human dependence on AI systems with reliance drills* (Version 5*).* arXiv. https://doi.org/10.48550/arXiv.2409.14055

Ibrahim, L., Collins, K. M., Kim, S. S. Y., Reuel, A., Lamparth, M., Feng, K., Ahmad, L., Soni, P., Kattan, A. E., Stein, M., Swaroop, S., Sucholutsky, I., Strait, A., Liao, Q. V., & Bhatt, U. (2025). *Measuring and mitigating overreliance is necessary for building human-compatible AI* (Version 1). arXiv. https://doi.org/10.48550/arXiv.2509.08010

Ioannou, A., Tavalaei, M. M., Vatn, D., & Mikalef, P. (2026). The role of explainability in AI-driven financial decision making. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-026-10727-1

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kahr, P. K., Rooks, G., Willemsen, M. C., & Snijders, C. C. P. (2024). Understanding trust and reliance development in AI advice: Assessing model accuracy, model explanations, and experiences from previous interactions. *ACM Transactions on Interactive Intelligent Systems, 14*(4), Article 29, 1–30. https://doi.org/10.1145/3686164

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2021). Trust in artificial intelligence: Meta-analytic findings. *Human Factors, 65*(2), 337–359. https://doi.org/10.1177/00187208211013988

Kazlaris, G. C., Keramopoulos, E., Bratsas, C., & Kokkonis, G. (2025). Augmented reality in education through collaborative learning: A systematic literature review. *Multimodal Technologies and Interaction, 9*(9), 94. https://doi.org/10.3390/mti9090094

Kelly, M., Kumar, A., Smyth, P., & Steyvers, M. (2023). Capturing humans' mental models of AI: An item response theory approach. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1723–1734. https://doi.org/10.1145/3593013.3594111

Klingbeil, A., Grützner, C., & Schreck, P. (2024. Trust and reliance on AI—An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior, 160*, 108352. https://doi.org/10.1016/j.chb.2024.108352

Know Center Research GmbH. (2026). *Data-driven immersive analytics in digital industries.* Retrieved April 10, 2026, from https://www.know-center.at/en/research/comet-modul/ddia/

Kornowicz, J., Pape, M., & Thommes, K. (2025). *Would I regret being different? The influence of social norms on attitudes toward AI usage* (Version 1). arXiv. https://doi.org/10.48550/arXiv.2509.04241

Kosch, T., Karolus, J., Zagermann, J., Reiterer, H., Schmidt, A., & Woźniak, P. W. (2023). A survey on measuring cognitive workload in human-computer interaction. *ACM Computing Surveys, 55*(13s), 1–39. https://doi.org/10.1145/3582272

Küper, A., Lodde, G. C., Livingstone, E., Schadendorf, D., & Krämer, N. (2025). Psychological factors influencing appropriate reliance on AI-enabled clinical decision support systems: Experimental web-based study among dermatologists. *Journal of Medical Internet Research, 27*, e58660. https://doi.org/10.2196/58660

Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., & Tan, C. (2021). *Towards a science of human-AI decision making: A survey of empirical studies* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2112.11471

Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 29–38. https://doi.org/10.1145/3287560.3287590

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Li, K., Wang, R., & Yang, C. (2025). Investigating the relationship and predictive modeling between mental workload and task performance in multitasking human–computer

interaction environments. *IEEE Access, 13*, 186189–186207. https://doi.org/10.1109/ACCESS.2025.3623798

Li, Y., Wu, B., Huang, Y., & Luan, S. (2024). Developing trustworthy artificial intelligence: Insights from research on interpersonal, human-automation, and human-AI trust. *Frontiers in Psychology, 15*. https://doi.org/10.3389/fpsyg.2024.1382693

Liang, G., Sloane, J. F., Donkin, C., & Newell, B. R. (2022). Adapting to the algorithm: How accuracy comparisons promote the use of a decision aid. *Cognitive Research: Principles and Implications, 7*(1). https://doi.org/10.1186/s41235-022-00364-y

Lu, Y., & Sarter, N. (2019). Eye tracking: A process-oriented method for inferring trust in automation as a function of priming and system reliability. *IEEE Transactions on Human-Machine Systems, 49*(6), 560–568. https://doi.org/10.1109/thms.2019.2930980

Lu, Z., Wang, D., & Yin, M. (2024). *Does more advice help? The effects of second opinions in AI-assisted decision making* (Version 1). arXiv. https://doi.org/10.48550/arXiv.2401.07058

Lu, Z., & Yin, M. (2021). Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3411764.3445562

Ma, S., Wang, X., Lei, Y., Shi, C., Yin, M., & Ma, X. (2024). *"Are you really sure?" Understanding the effects of human self-confidence calibration in AI-assisted decision making*. arXiv. https://doi.org/10.1145/3613904.3642671

Madritsch, J. H., Duricic, T., ElSayed, N., Kopeinik, S., & Veas, E. (2026). AI-powered conversational assistance in augmented reality for multi-step tasks. In *2026 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (pp. 519–529). IEEE. https://doi.org/10.1109/VR67842.2026.00072

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709. https://doi.org/10.2307/258792

McAllister, D. J. (1995). Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal, 38*(1), 24–59. https://doi.org/10.2307/256727

Microsoft. (2023, September 6). *MRTK3 overview*. Retrieved April 2, 2026, from https://learn.microsoft.com/en-us/windows/mixed-reality/mrtk-unity/mrtk3-overview/

Microsoft. (2025, December 18). *Holographic remoting player*. Retrieved April 2, from https://learn.microsoft.com/de-de/windows/mixed-reality/develop/native/holographic-remoting-player

Microsoft. (n.d.). *Microsoft HoloLens*. Retrieved April 2, 2026, from https://learn.microsoft.com/de-de/hololens/

Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other? In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 201–220). Erlbaum.

Noga, T., & Arnold, V. (2002). Do tax decision support systems affect the accuracy of tax compliance decisions? *International Journal of Accounting Information Systems, 3*(3), 125–144. https://doi.org/10.1016/S1467-0895(02)00034-9

OpenAI. (2026). *ChatGPT* (GPT-5.5) [Large language model]. Retrieved April 10, 2026, from https://chatgpt.com

OpenAI. (2026). *OpenAI API platform*. Retrieved April 2, 2026, from https://platform.openai.com/

OpenAI. (2026). *Text-to-speech guide*. Retrieved April 2, 2026, from https://developers.openai.com/api/docs/guides/text-to-speech

OpenAI. (2026). *Whisper*. Retrieved April 2, 2026, from https://openai.com/index/whisper/

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Perrig, S. A. C., Scharowski, N., & Brühlmann, F. (2023). Trust issues with trust scales: Examining the psychometric quality of trust measures in the context of AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–7). ACM. https://doi.org/10.1145/3544549.3585808

Pescetelli, N., Hauperich, A.-K., & Yeung, N. (2021). Confidence, advice seeking and changes of mind in decision making. *Cognition, 215*, 104810. https://doi.org/10.1016/j.cognition.2021.104810

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2021). *Manipulating and measuring model interpretability*. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Article 55, pp. 1–52). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445315

Pradhan, P., Rostami, M., Kamoonpuri, J., & Chung, J. (2023). The state of augmented reality in aerospace navigation and engineering. In *Applications of augmented reality—Current state of the art*. IntechOpen. https://doi.org/10.5772/intechopen.1002358

Pretty, E. J., & Guarese, R. (2025). *Mental workload wars: Audio-based XR awakens*. In *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp Companion '25)* (pp. 1–5). ACM. https://doi.org/10.1145/3714394.3756229

Rechkemmer, A., & Yin, M. (2022). When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3491102.3501967

Rejc, A. (2026). *Investigating trust in AI and overreliance on AI-generated guidance in an augmented reality environment* [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.20389976

Ries, A. J., Aroca-Ouellette, S., Roncone, A., & de Visser, E. J. (2025). Gaze-informed signatures of trust and collaboration in human-autonomy teams. *Computers in Human Behavior: Artificial Humans, 5*, 100171. https://doi.org/10.1016/j.chbah.2025.100171

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Romeo, G., & Conti, D. (2025). Exploring automation bias in human–AI collaboration: A review and implications for explainable AI. *AI & Society, 41*(1), 259–278. https://doi.org/10.1007/s00146-025-02422-7

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Introduction to special topic forum: Not so different after all: A cross-discipline view of trust. *Academy of Management Review, 23*(3), 393–404. https://www.jstor.org/stable/259285

Ryser, A., Allwein, F., & Schlippe, T. (2025). *Calibrated trust in dealing with LLM hallucinations: A qualitative study* (Version 1). arXiv. https://doi.org/10.48550/arXiv.2512.09088

Salimzadeh, S., & Gadiraju, U. (2024). When in doubt! Understanding the role of task characteristics on peer decision-making with AI assistance. *Proceedings of the 32nd*

ACM Conference on User Modeling, Adaptation and Personalization, 89–101. https://doi.org/10.1145/3627043.3659567

Semantic Scholar. (2026). *About Semantic Scholar*. Retrieved April 10, 2026, from https://www.semanticscholar.org/about

Shamszare, H., & Choudhury, A. (2023). Clinicians' Perceptions of Artificial Intelligence: Focus on Workload, Risk, Trust, Clinical Decision Making, and Clinical Integration. *Healthcare* (Basel, Switzerland), 11(16), 2308. https://doi.org/10.3390/healthcare11162308

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies, 146*, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Singh, S. (2025). Cognitive workload and interface performance: A neuroergonomic comparison of VR, AR, and traditional drone control systems. In *AHFE International* (Vol. 199). AHFE International. https://doi.org/10.54941/ahfe1006902

Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies, 52*(4), 701–717. https://doi.org/10.1006/ijhc.1999.0349

Steyvers, M., & Kumar, A. (2023). Three challenges for AI-assisted decision-making. *Perspectives on Psychological Science, 19*(5), 722–734. https://doi.org/10.1177/17456916231181102

Suzuki, Y., Wild, F., & Scanlon, E. (2023). Measuring cognitive load in augmented reality with physiological methods: A systematic review. *Journal of Computer Assisted Learning, 40*(2), 375–393. https://doi.org/10.1111/jcal.12882

Swaroop, S., Buçinca, Z., Gajos, K. Z., & Doshi-Velez, F. (2024). Accuracy-time tradeoffs in AI-assisted decision making under time pressure. *Proceedings of the 29th International Conference on Intelligent User Interfaces*, 138–154. https://doi.org/10.1145/3640543.3645206

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. Springer. https://doi.org/10.1007/978-1-4419-8126-4

Tang, Y., Situ, J., Cui, A. Y., Wu, M., & Huang, Y. (2025). LLM integration in extended reality: A comprehensive review of current trends, challenges, and future perspectives. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–24. https://doi.org/10.1145/3706598.3714224

Tao, D., Tan, H., Wang, H., Zhang, X., Qu, X., & Zhang, T. (2019). A systematic review of physiological measures of mental workload. *International Journal of Environmental Research and Public Health, 16*(15), 2716. https://doi.org/10.3390/ijerph16152716

Tejeda, H., Kumar, A., Smyth, P., & Steyvers, M. (2022). AI-assisted decision-making: A cognitive modeling approach to infer latent reliance strategies. *Computational Brain & Behavior, 5*(4), 491–508. https://doi.org/10.1007/s42113-022-00157-y

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185(4157):1124–1131.

Unity Technologies. (2024, June 12). *Unity 2022.3.33f1 release notes*. Retrieved April 2, 2026, from https://unity.com/de/releases/editor/whats-new/2022.3.33f1

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M., & Krishna, R. (2022*). Explanations can reduce overreliance on AI systems during decision-making* (Version 2). arXiv. https://doi.org/10.48550/arXiv.2212.06823

Vitti, M., Padovano, A., & Facchini, F. (2025). A review on cognitive workload for Industry 5.0. *Computers & Industrial Engineering, 207*, 111350. https://doi.org/10.1016/j.cie.2025.111350

Wang, X., Lu, Z., & Yin, M. (2022). Will you accept the AI recommendation? Predicting human behavior in AI-assisted decision making. *Proceedings of the ACM Web Conference 2022*, 1697–1708. https://doi.org/10.1145/3485447.3512240

Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., & Du, Z. (2024). *Artificial intelligence in education: A systematic literature review*. *Expert Systems with Applications, 252*, 124167. https://doi.org/10.1016/j.eswa.2024.124167

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention* (pp. 63–102). Academic Press.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50* (3), 449–455. https://doi.org/10.1518/001872008X288394

Wu, Z., Wang, Y., Langer, M., & Feit, A. M. (2025). RelEYEance: Gaze-based assessment of users' AI-reliance at run-time. *Proceedings of the ACM on Human-Computer Interaction, 9*(3), 1–18. https://doi.org/10.1145/3725841

Xu, G., Murthy, S. V., & Jia, B. (2025). Enhancing Intuitive Decision-Making and Reliance Through Human–AI Collaboration: A Review. *Informatics*, *12*(4), 135. https://doi.org/10.3390/informatics12040135

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300509

Yoo, S., Reza, S., Tarashiyoun, H., Ajikumar, A., & Moghaddam, M. (2024). *AI-integrated AR as an intelligent companion for industrial workers: A systematic review*. *IEEE Access, 12*, 191808–191827. https://doi.org/10.1109/access.2024.3516536

Zhang, R., Jin, X., Liu, M., & Tong, H. Y. (2025). The effectiveness of augmented reality/mixed reality in medical education: A meta-analysis. *BMC Medical Education, 25*(1). https://doi.org/10.1186/s12909-025-08166-8

Appendix A: Study 1 Measurement Instruments

Appendix A.1: Demographic Questionnaire

Participant Questionnaire

Please complete the following questionnaire. Your responses will remain confidential and will be used for research purposes only.

Age: _____

Native Language: _____

Gender

☐ Male

☐ Female

☐ Non-binary / Third gender

☐ Prefer not to say

How would you rate your proficiency in English?

☐ Beginner

☐ Intermediate

☐ Advanced

☐ Native or Bilingual

How familiar are you with the Juice Mixer?

☐ Not at all familiar

☐ Slightly familiar

☐ Moderately familiar

☐ Very familiar

☐ Extremely familiar

Do you have prior experience with Augmented Reality (AR) devices?

☐ None

☐ Tried once or twice

☐ Occasionally use AR devices

☐ Regular user of AR devices

Do you have prior experience with Augmented Reality (AR) assistants?

☐ None

☐ Tried once or twice

☐ Occasionally use AR devices

☐ Regular user of AR devices

Which AR devices have you used before? Select all that apply.

☐ Microsoft HoloLens

☐ Magic Leap

☐ Meta Quest (Oculus) with AR features

☐ Mobile AR Apps

☐ Other: _____

☐ None

How comfortable are you using AR devices?
1 = Not comfortable at all, 5 = Very comfortable

- _____

How familiar are you with AI or language models (e.g., ChatGPT, Siri, Alexa)?

☐ Not at all familiar

☐ Slightly familiar

☐ Moderately familiar

☐ Very familiar

☐ Extremely familiar

How comfortable are you interacting with AI or conversational agents?
1 = Not comfortable at all, 5 = Very comfortable

- _____

What is the highest degree or level of education you have completed?

☐ No formal education

☐ Primary school

☐ High school diploma or equivalent

☐ Bachelor's degree

☐ Master's degree

☐ Doctoral degree

☐ Other – please specify: _____

Appendix A.2: R-NASA TLX Administered After Each Task in Both Conditions in Study 1

The following are 20-point Likert scale questions. Please respond by marking the answer (number) that you think best describes your feelings, where 1 means *Very low* and 20 means *Very high.*

1. **Mental Demand**: How mentally demanding was the task?

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Very Low *Very High*

2. **Physical Demand**: How physically demanding was the task?

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Very Low *Very High*

3. **Temporal Demand**: How hurried or rushed was the pace of the task?

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Very Low *Very High*

4. **Performance**: How successful were you in accomplishing what you were asked to do?

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Very Low *Very High*

5. **Effort**: How hard did you have to work to accomplish your level of performance?

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Very Low *Very High*

6. **Frustration**: How insecure, discouraged, irritated, stressed and annoyed were you?

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Very Low *Very High*

Appendix A.3: Trust in AI Questionnaire Administered Before Engaging with Tasks in the AR+LLM Condition in Study 1

Please answer each question by selecting one response from the scale below. Mark your answer with an "X" in the box that best represents your response, where 1 means *Strongly disagree* and 5 means *Strongly Agree*.

1. I am confident in the AR system with AI support. I feel that it will work well.

1	2	3	4	5

Strongly disagree Strongly agree

2. The outputs of the AR system with AI will be very predictable.

1	2	3	4	5

Strongly disagree Strongly agree

3. The AR system with AI support will be very reliable. I can count on it to be correct all the time.

1	2	3	4	5

Strongly disagree Strongly agree

4. I feel safe that when I will rely on the AR system with AI support I will get the right answers.

1	2	3	4	5

Strongly disagree Strongly agree

5. The AR system with AI support will be efficient in that it will work very quickly.

1	2	3	4	5

Strongly disagree Strongly agree

6. I will like using the AR system with AI support for decision making.

1	2	3	4	5

Strongly disagree Strongly agree

Appendix A.4: Trust in AI Questionnaire Administered After Each Task in the AR+LLM Condition in Study 1

Please answer each question by selecting one response from the scale below. Mark your answer with an "X" in the box that best represents your response, where 1 means *Strongly disagree* and 5 means *Strongly Agree*.

1. I am confident in the AR system with AI support. I feel that it works well.

1	2	3	4	5
Strongly disagree				Strongly agree

2. The outputs of the AR system with AI are very predictable.

1	2	3	4	5
Strongly disagree				Strongly agree

3. The AR system with AI support is very reliable. I can count on it to be correct all the time.

1	2	3	4	5
Strongly disagree				Strongly agree

4. I feel safe that when I rely on the AR system with AI support I will get the right answers.

1	2	3	4	5
Strongly disagree				Strongly agree

5. The AR system with AI support is efficient in that it works very quickly.

1	2	3	4	5
Strongly disagree				Strongly agree

6. I like using the AR system with AI support for decision making.

1	2	3	4	5
Strongly disagree				Strongly agree

Appendix A.5: Trust in AI Questionnaire Administered Before Engaging with Tasks In the AR+PDF Condition in Study 1

Please answer each question by selecting one response from the scale below. Mark your answer with an "X" in the box that best represents your response, where 1 means *Strongly disagree* and 5 means *Strongly Agree*.

1. I am confident in the AR system with PDF support. I feel that it will work well.

1	2	3	4	5

Strongly disagree Strongly agree

2. The outputs of the AR system with PDF will be very predictable.

1	2	3	4	5

Strongly disagree Strongly agree

3. The AR system with PDF support will be very reliable. I can count on it to be correct all the time.

1	2	3	4	5

Strongly disagree Strongly agree

4. I feel safe that when I will rely on the AR system with PDF support I will get the right answers.

1	2	3	4	5

Strongly disagree Strongly agree

5. The AR system with PDF support will be efficient in that it will work very quickly.

1	2	3	4	5

Strongly disagree Strongly agree

6. I will like using the AR system with PDF support for decision making.

1	2	3	4	5

Strongly disagree Strongly agree

Appendix A.6: Trust in AI Questionnaire Administered After Each Task
In the AR+PDF Condition in Study 1

Please answer each question by selecting one response from the scale below. Mark your answer with an "X" in the box that best represents your response, where 1 means *Strongly disagree* and 5 means *Strongly Agree*.

1. I am confident in the AR system with PDF support. I feel that it works well.

1	2	3	4	5
Strongly disagree				Strongly agree

2. The outputs of the AR system with PDF are very predictable.

1	2	3	4	5
Strongly disagree				Strongly agree

3. The AR system with PDF support is very reliable. I can count on it to be correct all the time.

1	2	3	4	5
Strongly disagree				Strongly agree

4. I feel safe that when I rely on the AR system with PDF support I will get the right answers.

1	2	3	4	5
Strongly disagree				Strongly agree

5. The AR system with PDF support is efficient in that it works very quickly.

1	2	3	4	5
Strongly disagree				Strongly agree

6. I like using the AR system with PDF support for decision making.

1	2	3	4	5
Strongly disagree				Strongly agree

Appendix B: Study 2 Measurement Instruments

Appendix B.1: Demographic Questionnaire Administered in Study 2

Participant Questionnaire

Please complete the following questionnaire. Your responses will remain confidential and will be used for research purposes only.

Age: _____

Native Language: _____

Gender

☐ Male

☐ Female

☐ Non-binary / Third gender

☐ Prefer not to say

How would you rate your proficiency in English?

☐ Beginner

☐ Intermediate

☐ Advanced

☐ Native or Bilingual

How familiar are you with the Juice Mixer?

☐ Not at all familiar

☐ Slightly familiar

☐ Moderately familiar

☐ Very familiar

☐ Extremely familiar

Do you have prior experience with Augmented Reality (AR) devices?

☐ None

☐ Tried once or twice

☐ Occasionally use AR devices

☐ Regular user of AR devices

Do you have prior experience with Augmented Reality (AR) assistants?

☐ None

☐ Tried once or twice

☐ Occasionally use AR devices

☐ Regular user of AR devices

Which AR devices have you used before? Select all that apply.

☐ Microsoft HoloLens

☐ Magic Leap

☐ Meta Quest (Oculus) with AR features

☐ Mobile AR Apps

☐ Other: _____

☐ None

How comfortable are you using AR devices?
1 = Not comfortable at all, 5 = Very comfortable

- _____

How familiar are you with AI or language models (e.g., ChatGPT, Siri, Alexa)?

☐ Not at all familiar

☐ Slightly familiar

☐ Moderately familiar

☐ Very familiar

☐ Extremely familiar

How comfortable are you interacting with AI or conversational agents?
1 = Not comfortable at all, 5 = Very comfortable

- _____

What is the highest degree or level of education you have completed?

☐ No formal education

☐ Primary school

☐ High school diploma or equivalent

☐ Bachelor's degree

☐ Master's degree

☐ Doctoral degree

☐ Other – please specify: _____

Appendix B.2: R-NASA TLX Questionnaire Administered After Each Task in Study 2

The following are 20-point Likert scale questions. Please respond by marking the answer (number) that you think best describes your feelings, where 1 means *Very low* and 20 means *Very high.*

1. **Mental Demand**: How mentally demanding was the task?

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Very Low *Very High*

2. **Physical Demand**: How physically demanding was the task?

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Very Low *Very High*

3. **Temporal Demand**: How hurried or rushed was the pace of the task?

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Very Low *Very High*

4. **Performance**: How successful were you in accomplishing what you were asked to do?

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Very Low *Very High*

5. **Effort**: How hard did you have to work to accomplish your level of performance?

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Very Low *Very High*

6. **Frustration**: How insecure, discouraged, irritated, stressed and annoyed were you?

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Very Low *Very High*

Appendix B.3: Trust in AI Questionnaire Administered Before Engaging with Tasks in Study 2

Please answer each question by selecting one response from the scale below. Mark your answer with an 'X' in the box that best represents your response, where 1 means *Strongly disagree*, 2 means *Disagree*, 3 *Neither agree nor disagree*, 4 means *Agree*, and 5 means *Strongly agree.*

1. I am confident in the AI. I feel that it will work well.

1	2	3	4	5

Strongly disagree Strongly agree

2. The outputs of the AI will be very predictable.

1	2	3	4	5

Strongly disagree Strongly agree

3. The AI will be very reliable. I can count on it to be correct all the time.

1	2	3	4	5

Strongly disagree Strongly agree

4. I feel safe that when I will rely on the AI I will get the right answers.

1	2	3	4	5

Strongly disagree Strongly agree

5. The AI will be efficient in that it will work very quickly.

1	2	3	4	5

Strongly disagree Strongly agree

6. I will like using the AI for decision making.

1	2	3	4	5

Strongly disagree Strongly agree

Appendix B.4: Trust in AI Questionnaire Administered After Each Task in Study 2

Please answer each question by selecting one response from the scale below. Mark your answer with an 'X' in the box that best represents your response, where 1 means *Strongly disagree*, 2 means *Disagree*, 3 *Neither agree nor disagree*, 4 means *Agree*, and 5 means *Strongly agree.*

1. I am confident in the AI. I feel that it works well.

1	2	3	4	5
Strongly disagree				Strongly agree

2. The outputs of the AI are very predictable.

1	2	3	4	5
Strongly disagree				Strongly agree

3. The AI is very reliable. I can count on it to be correct all the time.

1	2	3	4	5
Strongly disagree				Strongly agree

4. I feel safe that when I rely on the AI I will get the right answers.

1	2	3	4	5
Strongly disagree				Strongly agree

5. The AI is efficient in that it works very quickly.

1	2	3	4	5
Strongly disagree				Strongly agree

6. I like using the AI for decision making.

1	2	3	4	5
Strongly disagree				Strongly agree

Appendix B.5: Trust in LLM Index Administered Before Engaging with Tasks in Study 2

Please answer each question by selecting one response from the scale below. Mark your answer with an 'X' in the box that best represents your response, where 1 means *Strongly disagree*, 2 means *Disagree*, 3 *Neither agree nor disagree*, 4 means *Agree*, and 5 means *Strongly agree*.

1. I would feel a sense of dismay if my interactions with an LLM were suddenly disrupted or halted.

1	2	3	4	5

Strongly disagree Strongly agree

2. If I share my concerns with LLMs, I know these agents will respond constructively and caringly.

1	2	3	4	5

Strongly disagree Strongly agree

3. I invest plenty of time developing and improving my prompts to interact with LLMs.

1	2	3	4	5

Strongly disagree Strongly agree

4. I can rely on LLMs not to make my job more difficult by careless work.

1	2	3	4	5

Strongly disagree Strongly agree

5. Despite trusting LLMs' results overall, the last word is always mine.

1	2	3	4	5

Strongly disagree Strongly agree

6. I tend to trust LLMs more than other people.

1	2	3	4	5

Strongly disagree Strongly agree

Appendix B.6: Trust in LLM Index Administered After Each Task in Study 2

Please answer each question by selecting one response from the scale below. Mark your answer with an 'X' in the box that best represents your response, where 1 means *Strongly disagree*, 2 means *Disagree*, 3 *Neither agree nor disagree*, 4 means *Agree*, and 5 means *Strongly agree*.

1. I would feel a sense of dismay if my interactions with an LLM were suddenly disrupted or halted.

1	2	3	4	5

Strongly disagree Strongly agree

2. If I share my concerns with LLMs, I know these agents will respond constructively and caringly.

1	2	3	4	5

Strongly disagree Strongly agree

3. I invest plenty of time developing and improving my prompts to interact with LLMs.

1	2	3	4	5

Strongly disagree Strongly agree

4. I can rely on LLMs not to make my job more difficult by careless work.

1	2	3	4	5

Strongly disagree Strongly agree

5. Despite trusting LLMs' results overall, the last word is always mine.

1	2	3	4	5

Strongly disagree Strongly agree

6. I tend to trust more LLMs than other people.

1	2	3	4	5

Strongly disagree Strongly agree

Appendix C: Overview of Correct and Erroneous LLM-Based Instructions in Task A

Table D1

Overview of Correct and Erroneous LLM-Based Instructions in Task A

Step	Correct Instructions	Erroneous instructions
4.2.1 Step 1	Fill the smaller vessel with the provided red water. The **larger** vessel should be filled with only water. The fill values are shown in Figure 4.1.	Fill the smaller vessel with the provided red water. The **smaller** vessel should be filled with only water. The fill values are shown in Figure 4.1.
4.2.3 Step 3	**Place the lid firmly onto the vessels. Insert the temperature and pH sensors into their respective ports until they are fully seated to ensure accurate readings.** Figure 4.3 shows the correct placement of the pH sensor and the temperature sensor. Please perform this procedure for both vessels	**Insert the temperature and pH sensors into the vessels. Once both sensors are inserted, place the lid firmly on top of the sensors** to ensure correct readings. Figure 4.3 shows the correct placement of the pH sensor and the temperature sensor. Please perform this procedure for both vessels.
4.2.5 Step 5	Juice A contains equal portions of both vessels. Activate the pump using the corresponding switch. The switch can be turned to the left for continuous operation, allowing hands-free use, or toggled to the right for short-term activation, in which case you must hold the switch until the mixing is complete. Use the same switch to turn the pump off. To pump the water, please use the switch for **pump 4**, for the other vessel the switch for pump 6. Fill the result vessel to a height of approximately 2-3 cm.	Juice A contains equal portions of both vessels. Activate the pump using the corresponding switch. Activate one pump after another. The switch can be turned to the left for continuous operation, allowing hands-free use, or toggled to the right for short-term activation, in which case you must hold the switch until the mixing is complete. Use the same switch to turn the pump off. To pump the water, please use the switch for **pump 3**, for the other vessel, use the switch for pump 6. Fill the result vessel to a height of approximately 2-3 cm.

Note. The *Correct Instructions* column presents the instructions provided in the PDF documentation, as displayed on the PDF display. The *Erroneous Instructions* column presents the manipulated instructions used as the knowledge source for the LLM-based assistant, where bold text indicates the introduced errors that formed the basis of the LLM-generated outputs.

Appendix D: Ethics Committee Approval Document

Graz University of Technology

The Chairman of the TU Graz Ethics Committee

Univ.-Prof. Mag.phil. Dr.theol.
Thomas Gremsl

thomas.gremsl@uni-graz.at

Office:

Rechbauerstraße 12
A-8010 Graz
Tel.: +43 (0)316 873 6003

ethikkommission@tugraz.at

Mr.
Priv.-Doz. Dipl.-Ing. Dr.techn. Dominik Kowald
[redacted]
Sandgasse 34/2
8010 Graz
Austria

Graz, 03.11.2025

GZ: EK-92/2025

VOTE

Ethics Committee

On the basis of the documents submitted, the Ethics Committee of Graz University of Technology has decided that the research project „Exploring User Behavior and Task Performance in AI-Supported Digital Environments" is ethically acceptable.

The vote of the Ethics Committee in no way affects the sole responsibility of the researchers for the proper conduct of the study in compliance with all relevant legal provisions and guidelines.

Procedure

Initial contact with the Ethics Committee's office on 15.10.2025.

Study description, catalog of criteria and informed consent on am 21.10.2025.

Checked under data protection law by Ms. Waxnegger, Know Center Research GmbH's legal counsel.

Meeting of the Ethics Committee on 03.11.2025.

Decision of the Ethics Committee on 03.11.2025.

[redacted]

November 2025 *Ethics Committee / Office*

Appendix E: Supplementary Statistical Analyses Tables for Study 1

Table E1

Descriptive Statistics for Gaze Duration on the PDF Display Across Tasks in the AR+PDF Condition (n = 18)

Condition	AR+PDF		
Display	PDF display		
Task	*M*	*SD*	*Mdn*
A	280,789.83	98,265.41	245,880.50
B	310,062.94	92,680.71	320,566.50
C	139,103.56	45,981.94	146,571.50

Note. M = mean; *SD* = standard deviation; *Mdn* = median. Gaze duration values are reported in milliseconds. Higher values indicate longer visual attention allocated to the respective display area during task completion in the *AR+PDF* condition.

Table E2

Descriptive Statistics for Gaze Duration on Chat Window and PDF Display Across Tasks in the AR+LLM Condition (n = 18)

Condition	AR+LLM					
Display	PDF display			Chat Window		
Task	*M*	*SD*	*Mdn*	*M*	*SD*	*Mdn*
A	150,827.22	78,779.82	146,844	223,917.72	91,892.42	230,382.50
B	180,183.28	78,189.60	169,281.50	253,954.67	106,929.60	231,938.50
C	99,472.39	41,019.85	89,285	176,073.44	89,301.22	172,450.50

Note. M = mean; *SD* = standard deviation; *Mdn* = median. Gaze duration values are reported in milliseconds. Higher values indicate longer visual attention allocated to the respective display area during task completion in the *AR+LLM* condition.

Table E3

Spearman and Pearson Correlations Between Trust in AI, Mental Workload and Gaze Duration on PDF Display and Chat Window Across Tasks in AR+LLM Condition (n = 18)

Task	Variables	*r*	*p*	*ρ*	*p*
A	Trust in AI – PDF gaze	−.08	.76	−.06	.8
	Trust in AI – Chat gaze	.18	.49	.26	.30
	R-TLX – PDF gaze	.41	.09	.35	.16
	R-TLX – Chat gaze	.01	.96	−.14	.57
B	Trust in AI – PDF gaze	−.33	.18	−.20	.43
	Trust in AI – Chat gaze	.05	.84	.09	.73
	R-TLX – PDF gaze	.42	.08	.32	.20
	R-TLX – Chat gaze	.33	.18	.27	.28
C	Trust in AI – PDF gaze	−.29	.25	−.23	.35
	Trust in AI – Chat gaze	.27	.29	.28	.26
	R-TLX – PDF gaze	.15	.55	.12	.62
	R-TLX – Chat gaze	−.18	.48	−.15	.57

Note. R-TLX = Raw Task Load Index. *r* = Pearson correlation coefficient; *ρ* = Spearman's rank correlation coefficient. Correlations were computed separately for each task within the *AR+LLM* condition. * $p < .05$, ** $p < .01$, *** $p < .001$.

Appendix F: Supplementary Statistical Analyses Tables for Study 2

Table F1

Descriptive Statistics of Trust in AI Results Across Groups and Tasks

Group	Error (n = 11)			No Error (n = 7)		
Task	*M*	*SD*	*Mdn*	*M*	*SD*	*Mdn*
Baseline	3.14	0.5	3.17	2.98	0.49	2.67
Task A	2.92	0.83	3.33	2.17	0.45	2.17
Task B	2.86	0.83	2.67	2.19	0.78	2
Task C	2.86	0.87	2.67	2.17	0.63	2.5

Note. *M* = mean; *SD* = standard deviation; *Mdn* = median. Trust in AI scores were calculated separately for the *Error* and *No Error* groups across the baseline and experimental tasks. Higher scores indicate greater trust in the AI system.

Table F2

Descriptive Statistics of Trust in LLM Results Across Groups and Tasks

Group	Error (n = 11)			No Error (n = 7)		
Task	*M*	*SD*	*Mdn*	*M*	*SD*	*Mdn*
Baseline	2.89	0.74	3.17	3.12	0.74	2.83
Task A	2.82	0.75	2.83	2.83	0.54	2.67
Task B	2.95	0.66	3	2.79	0.66	2.67
Task C	2.89	0.75	3.17	2.86	0.56	2.83

Note. *M* = mean; *SD* = standard deviation; *Mdn* = median. Trust in LLM scores were calculated separately for the *Error* and *No Error* groups across the baseline and experimental tasks. Higher scores indicate greater trust in the LLM.

Table F3

Mixed ANOVA Results for Trust in LLM

Effect	*df*	*F*	*p*	$\eta^2_{(G)}$
Group	1, 16	0.00	.98	< .001
Task	3, 48	0.95	.43	.01
Group × Task	3, 48	1.04	.38	.01

Note. $\eta^2_{(G)}$ = generalized eta squared. Greenhouse–Geisser corrections were applied where the assumption of sphericity was violated. No significant effects were observed.

Table F4

Descriptive Statistics of Task Performance Rates (in %) Across Groups and Tasks

Group	Error (n = 11)			No Error (n = 7)		
Task	*M*	*SD*	*Mdn*	*M*	*SD*	*Mdn*
A	69.32	15.04	68.75	94.2	8.54	100
B	88.89	8.87	88.89	96.03	5.28	100
C	76.99	11.46	78.13	87.5	8.84	93.75

Note. M = mean; *SD* = standard deviation; *Mdn* = median. Task Performance rates were calculated separately for the *Error* and *No Error* groups across the experimental tasks A–C. Higher values indicate better task performance.

Table F5

Overdispersion Diagnostics for Poisson Regression Models

Model	Pearson χ^2	*df*	Ratio
Trust in AI	17.35	16	1.08
Trust in LLM	21.03	16	1.31
Psychological predictors	9.89	12	0.82
Experience predictors	18.18	15	1.21
Full model	8.45	10	0.85

Note. Ratio = Pearson χ^2 divided by degrees of freedom (*df*). Values close to 1 indicate no substantial overdispersion in the Poisson regression models.

Table F6

Poisson Regression Models Predicting Overreliance Frequency

Model	Predictor	*b*	*SE*	*z*	*p*	*IRR*	95% CI (IRR)
Trust AI only	Trust in AI	0.8	0.31	2.58	.01	2.22	[1.21, 4.08]
Trust LLM only	Trust in LLM	0.16	0.35	0.46	.64	1.18	[0.59, 2.33]
Psychological predictors	Trust in AI	1.19	0.59	2.03	.04	3.3	[1.04, 10.46]
	Trust in LLM	–0.76	0.55	–1.38	.17	0.47	[0.16, 1.38]
	R- TLX	0.12	0.11	1.11	.27	1.12	[0.91, 1.38]
	AI comfortability	0.5	0.31	1.61	.11	1.65	[0.90, 3.04]
	AR comfortability	0	0.27	0	.999	1	[0.59, 1.69]
Experience predictors	AR experience	–0.47	0.4	–1.16	.25	0.63	[0.29, 1.38]
	AI familiarity	0.59	0.31	1.89	.06	1.81	[0.98, 3.35]
Full model	Trust in AI	1.41	0.84	1.68	.09	4.11	[0.79, 21.43]
	Trust in LLM	–0.67	0.72	–0.94	.35	0.51	[0.13, 2.08]

R-TLX	0.19	0.18	1.05	.295	1.21	[0.85, 1.72]
AI comfortability	0.08	0.56	0.14	.89	1.08	[0.36, 3.26]
AR comfortability	0.14	0.28	0.5	.62	1.15	[0.66, 2.01]
AR experience	–0.97	0.8	–1.2	.23	0.38	[0.08, 1.84]
AI familiarity	0.83	0.59	1.41	.16	2.29	[0.72, 7.30]

Note. b = unstandardized coefficient; *SE* = standard error; *IRR* = incidence rate ratio; CI = confidence interval. Positive coefficients indicate higher predicted overreliance frequency. Confidence intervals are reported for IRRs. *$p < .05$.

Table F7

Descriptive Statistics of Gaze Behavior Results per Display Across Groups and Tasks (in ms)

Display	PDF display					
Group	Error (n = 11)			No Error (n = 7)		
Task	*M*	*SD*	*Mdn*	*M*	*SD*	*Mdn*
Task A	333,322.70	209,831.60	312,301	289,163.70	86,780.01	272,612
Task B	195,660.50	75,666.69	190,412	199,804.30	63,382.82	192,715
Task C	138,324.10	44,421.32	134,092	153,852.30	27,318.49	148,715

Display	Chat window					
Group	Error (n = 11)			No Error (n = 7)		
Task	*M*	*SD*	*Mdn*	*M*	*SD*	*Mdn*
Task A	364,600.40	112,281.70	362,314	290,628.00	180,664.20	258,318
Task B	236,248.40	103,476.10	219,501	157,338.60	62,969.38	161,942
Task C	178,619.60	70,974.83	178,515	135,355.40	44,082.21	121,577

Note. *M* = mean; *SD* = standard deviation; *Mdn* = median. Values represent gaze duration on the PDF display and chat window (in ms), reported separately for each display and group (*Error, No Error*) and across tasks A to C, with higher values indicating longer gaze duration.

Table F8

Descriptive Statistics of Mental workload (R-TLX) Scores Across Groups and Tasks

Group	Error (n = 11)			No Error (n = 7)		
Task	*M*	*SD*	*Mdn*	*M*	*SD*	*Mdn*
A	8.48	3.37	9	8.12	3.38	9.67
B	6.39	3.86	6	6	3.15	5.67
C	6.35	3.96	8	5.55	3.62	4

Note. *M* = mean; *SD* = standard deviation; *Mdn* = median. Cognitive workload scores were calculated separately for the *Error* and *No Error* groups across the tasks. Higher scores indicate greater perceived mental workload.

Table F9

Correlations Between Trust Measures and Gaze Behavior Across Tasks (N = 18 per task)

Task	A				B				C			
Variables	*r*	*p*	*ρ*	*p*	*r*	*p*	*ρ*	*p*	*r*	*p*	*ρ*	*p*
Trust in AI – PDF	−.33	.18	−.26	.31	−.42	.08	−.36	.14	−.44	.07	−.35	.15
Trust in AI – Chat	.34	.17	.17	.50	.42	.08	.35	.15	**.57***	.01	**.47***	.05
Trust in LLM – PDF	−.12	.63	−.03	.90	−.42	.08	−.43	.08	−.14	.58	−.11	.67
Trust in LLM – Chat	.18	.48	.26	.29	.22	.38	.22	.39	.18	.47	.27	.27

Note. r = Pearson correlation coefficient; *ρ* = Spearman's rank correlation coefficient. PDF and Chat refer to gaze duration on the PDF display and AI chat window, respectively. Positive correlations indicate that higher trust in AI was associated with longer gaze duration on Chat window. *$p < .05$.

Table F10

Correlations Between Mental Workload and Gaze Behavior Across Displays and Tasks (n = 18 per task)

Task	Display	*ρ*	*p*	95% CI
A	Chat window	−.13	.597	[−.57, .36]
	PDF display	0.13	.62	[−.36, .56]
B	Chat window	0.13	.60	[−.36, .56]
	PDF display	0.18	.47	[−.31, .60]
C	Chat window	−.03	.89	[−.49, .44]
	PDF display	0.01	.96	[−.46, .48]

Note. ρ = Spearman's rank correlation coefficient; *CI* = confidence interval. Chat window and PDF display refer to gaze duration on the respective display areas during task completion. Confidence intervals are reported at the 95% level. No significant results were observed.

Table F11

Mixed-Design ANOVA Results for Mental Workload (R-TLX Scores)

Effect	*df*	*F*	*p*	η^2G
Group	1, 16	0.11	0.74	0.006
Task	2, 32	**6.67***	.004	0.08
Group × Task	2, 32	0.06	0.94	< .001

Note. *$p < .05$; η^2G = generalized eta squared. *$p < .05$.